Regressão "Ridge":
Um Método Alternativo para o Mal Condicionamento da Matriz das Regressoras

Cristiane Reynaldo

Orientador: Prof. Dr. Reinaldo Charnet

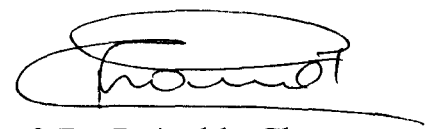
Instituto de Matemática, Estatística e Computação Científica, UNICAMP

Nov-1997

Regressão "Ridge":
Um Método Alternativo para o Mal Condicionamento da Matriz das Regressoras

Este exemplar corresponde a redação final da dissertação devidamente corrigida e defendida por Cristiane Reynaldo e aprovada pela Comissão Julgadora.

Campinas, 06 de novembro de 1997



Prof. Dr. Reinaldo Charnet
Orientador.

Dissertação apresentada ao Instituto de Matemática, Estatística e Computação Científica, UNICAMP, como requisito parcial para obtenção do Título de Mestre em Estatística.

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO IMECC DA UNICAMP

Reynaldo, Cristiane

R331r Regressão "Ridge": um método alternativo para o mal condicionamento da matriz das regressoras / Cristiane Reynaldo -- Campinas, [S.P. :s.n.], 1997.

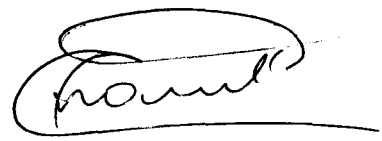
Orientador : Reinaldo Charnet

Dissertação (mestrado) - Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Multicolinearidade. 2. .Vício. I. Charnet, Reinaldo. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Dissertação de Mestrado defendida e aprovada em 06 de novembro de 1997

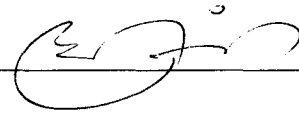
pela Banca Examinadora composta pelos Profs. Drs.



Prof (a). Dr (a). REINALDO CHARNET

Clarice Azevedo de Luna Freire

Prof (a). Dr (a). CLARICE AZEVEDO DE LUNA FREIRE



Prof (a). Dr (a). JOSÉ ANTONIO CORDEIRO

Dedico este trabalho

a meus pais.

Agradecimentos

A Deus, por tudo.

Aos meu pais, Milton e Angelina, pelo amor, carinho, apoio financeiro e por diversas vezes abdicarem de seus sonhos a favor dos meus.

Aos meus irmãos, Luciane e Renato, pela incrível amizade e por, simplesmente, existirem.

Em especial, ao Rogério pelo carinho, apoio e por grandes momentos.

Ao meu orientador, Reinaldo, pela paciência, amizade, e orientação que foi fundamental para o desenvolvimento deste trabalho.

A Verônica, Rosi, Paula, Raquel, Lianca e todas outras da Carmen's house, as que permanecem e as que já se foram, pela amizade, medicações e os almoços de finais de semana.

A Daniela, Fernando, novamente, Luciane e Odail pelas valiosas sugestões e fornecimento dos dados.

A todos meus amigos da estatística, principalmente, os da turma de 95 que foram verdadeiros cúmplices durante todo meu mestrado.

Aos professores e funcionários do IMECC, pela formação e ajudas recebidas.

A CAPES e Funcamp pelo apoio financeiro.

A turma do vôlei e tantas outras pessoas que marcaram um período inesquecível de minha vida.

o meu, muito obrigada.

"Deus

Dai-me serenidade para aceitar as coisas que não posso mudar, coragem para mudar as que posso e sabedoria para perceber a diferença"

Santo Agostinho.

Sumário

Resumo

Nas análises de regressão linear múltipla existem muitas situações onde o mal condicionamento da matriz das regressoras está presente. De forma geral, o que se costuma fazer é eliminar uma das variáveis do modelo de regressão. Entretanto, supomos que este processo já foi realizado e o mal condicionamento ainda permanece.

Essa situações não é ilusória, uma vez que existem muitos exemplos em dados econômicos.

Assim, sugerimos a regressão "ridge" como um método alternativo. Existem várias maneiras de se obter os estimadores "ridge", aqui, fornecemos algumas delas.

Portanto, o objetivo deste trabalho é comparar os estimadores "ridge" e mostrar suas vantagens sobre os estimadores de mínimos quadrados, quando os dados estão mal condicionados.

Introdução

Este trabalho é um estudo comparativo entre os estimadores de mínimos quadrados, os estimadores "ridge" e entre os métodos de se obter este último.

Para tal, deveremos ter o conhecimento do método estatístico que relaciona duas variáveis. Segundo Karl Pearson[1], quem, pela primeira vez relatou a existência de correlação foi August Bravais, em 1846. Mais tarde, Sir Francis Galton,1877, em seu livro com o título: *Typical Laws of Heredity in Man*, estudava a semelhança das crianças com seus pais. Muitos estatísticos ficaram fascinados por esta questão e uniram imensos conjuntos de dados em busca de uma resposta. Karl Pearson (1857 - 1936) um dos seus discípulos, durante várias gerações estudou a semelhança entre os membros das famílias, medindo a altura de 1078 pais e seus filhos na adolescência, um filho por pai. Esta lista de 1078 pares de alturas era impossível de analisar, mas a relação entre as duas variáveis foi possível, representando-as em um gráfico chamado *diagrama de dispersão.* Assim, pôde observar que os filhos de pais baixos eram menores que a média dos pais mas não tão baixos como o menor destes e os filhos de pais altos eram maiores que média dos pais, mas não tão alto como o maior dos pais. Assim, aparece o simbolo *r* da terminologia *reversão* e que somente mais tarde Galton determina *regressão*. O *r* corresponde ao nosso símbolo de coeficiente de correlação a qual primeiramente foi denominado de *reversão* e não de *regressão*.

Hoje, sabemos que a análise de *Modelos de Regressão* faz parte de um estudo muito mais amplo, incluindo várias especificações. Aqui, nos deteremos à análise de modelos de regressão linear[2].

Supondo o conhecimento do leitor sobre regressão linear simples, iniciaremos o primeiro capítulo com um caso mais geral, a *Regressão Linear Múltipla*. Nele veremos estimação dos parâmetros, propriedades dos estimadores, estimação da variância, análise da variância, o coeficiente de determinação, o porquê e quais as vantagens de padronizarmos os conjuntos de dados e, por último, a decomposição de valores singulares. Todos esses tópicos são essenciais para o aprendizado da regressão "ridge".

---

[1] ver bibliografia.

[2] A linearidade referida está relacionada com os parâmetros.

Agora, o que é regressão "ridge"? Essa é uma pergunta que será respondida no capítulo 3 deste trabalho. Antes disso, o leitor deve saber que a regressão "ridge" é utilizada quando os dados apresentam-se com multicolinearidade aproximada, também conhecida na literatura por mal condicionamento.

Assim, no segundo capítulo veremos o que é *Multicolinearidade*, seus efeitos, como detectá-la e as soluções apresentadas pelos pesquisadores.

Com isso, chegamos ao terceiro capítulo com a *Regressão "Ridge"*, nele apresentaremos a forma geral do estimador "ridge", suas propriedades, as medidas de comparação: erro quadrático médio total dos estimadores (EQMT) e erro quadrático médio total do predito (EQMTP), veremos os teoremas que avaliam estas medidas e os métodos para obtermos os estimadores "ridge".

Os leitores mais esclarecidos, neste assunto, podem estar se perguntando: por que não eliminarmos algumas variáveis do modelo e aplicarmos a regressão linear múltipla?, um método que encontramos implementado em vários pacotes estatísticos onde só precisamos adicionar os dados e todos os cálculos é feito em poucos segundos.

Nossa justificativa para tal é que supomos ter eliminado todas as variáveis possíveis do modelo e mesmo assim o mal condicionamento, ainda, permanece. Consideramos inviável a exclusão de qualquer outra variável, pois isto acarretaria em muita perda de informação. Estas suposições não são ilusórias ou nocivas, uma vez que muitas destas situações encontramos em dados econômicos. Entretanto, são nestas ocasiões que sugerimos a regressão "ridge" como um método alternativo para se obter os estimadores.

Entendidas estas suposições e vista a parte teórica, no capítulo 4 todas essas informações serão utilizadas aplicando na simulação e num exemplo com dados reais. Neste capítulo, consideraremos a *Simulação* de vários conjuntos de dados com correlações pré determinadas, que induzem ao mal condicionamento, assim, encontraremos os estimadores de mínimos quadrados e os estimadores "ridge", fornecidos no terceiro capítulo, e compara-lo-emos baseados no EQMT, na variância, no vício e no EQMTP. Por último, utilizaremos desses resultados escolhendo alguns métodos e aplicaremos no conjunto de dados não simulado.

Portanto, este será nosso objetivo. Mostrar as possíveis vantagens dos métodos "ridge" sobre os mínimos quadrados e quais dos métodos propostos se apresentam melhor, comparando-os baseados nas medidas citadas acima.

Capítulo 1

Regressão Linear Múltipla

Sabemos que há uma aplicabilidade muito ampla nos modelos lineares de regressão com várias especificações de modelos, porém não serão citados aqui. Nos deteremos apenas ao modelo de regressão linear múltipla.

Contudo, nosso objetivo é desenvolver os conceitos deste modelo de regressão, que serão utilizados, ao longo deste trabalho. Com essa finalidade relacionamos uma variável dependente, y, com outras variáveis independentes,(x_1,x_2, ...,x_p), ajustando assim, um modelo ao conjunto de dados disponíveis.

Baseados neste modelo estimamos os parâmetros através do método dos mínimos quadrados, estudaremos suas propriedades e faremos a análise da variância.

Procuramos conduzir este capítulo de forma clara e objetiva para que o leitor tenha uma boa compreensão dos próximos capítulos, e sendo assim, adicionamos alguns conceitos matemáticos que serão necessários para o desenvolvimento deste e por último, incluimos, *decomposição de valores singulares*, que será utilizada em todo trabalho.

1.1 Modelo Regressão

Análise de Regressão é uma técnica estatística para investigar e modelar a relação entre variáveis. Essa relação pode ser linear ou não-linear. Quando for linear entre as variáveis será denominada modelo de regressão linear. Este, em sua forma matricial será representado por:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \qquad (1.1)$$

onde, para uma amostra de tamanho n, teremos:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \text{ vetor de variáveis respostas}$$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \text{ matriz de variáveis regressoras}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \text{ vetor de parâmetros desconhecidos}$$

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \text{ vetor de componentes aleatórias.}$$

1.1.1 Estimação dos Parâmetros

Antes de realizarmos a estimação do vetor β, iniciaremos com algumas definições, envolvendo teoria das matrizes (ver *Graybill,* 1983).

Definição 1: O posto de uma matriz **A** (n × p) será dado pela maior ordem possível das submatrizes quadradas de A, com determinante diferente de zero.

Definição 2: Uma matriz quadrada **A** (p × p) será dita não-singular se seu posto for p, quando se dirá que a matriz **A** tem posto completo. Neste caso existirá uma única matriz $\mathbf{A}^{-1}$ tal que :

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}.$$

Uma implicação imediata destas definições é a seguinte: se **A** é não-singular então, seu determinante será não nulo e existirá uma única matriz $\mathbf{A}^{-1}$ inversa de **A.**

Existem muitos métodos para obter-se a estimativa dos parâmetros no modelo. Discutiremos, aqui, o chamado quadrados mínimos uma vez que este é um dos mais utilizados na literatura. As estimativas dos parâmetros, usando os métodos de quadrados mínimos, são encontrados de forma que minimize a soma dos quadrados do erro, ou seja, $\varepsilon^T\varepsilon = \phi(\beta)$.

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \Rightarrow \varepsilon = \mathbf{y} - \mathbf{X}\beta$$

$$\varepsilon^T\varepsilon = \phi(\beta) = (\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta) = \mathbf{y}^T\mathbf{y} - 2\beta^T(\mathbf{X}^T\mathbf{y}) + \beta^T\mathbf{X}^T\mathbf{X}\beta$$

$$\frac{\partial\phi(\beta)}{\partial\beta} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\beta$$

igualando a zero, obteremos o seguinte sistema, chamado equações normais

$$\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{y} \qquad (1.2)$$

onde **b** é estimador de β.

A solução do sistema (1.2) está diretamente relacionada com a estrutura da matriz $\mathbf{X}^T\mathbf{X}$. De forma que, se $\mathbf{X}^T\mathbf{X}$ for não-singular, (1.2) terá uma única solução e será igual ao estimador:

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Por outro lado, se $\mathbf{X}^T\mathbf{X}$ for singular, (1.2) poderá ser resolvida usando-se inversa generalizada, mas não terá solução única. Caso $\mathbf{X}^T\mathbf{X}$ seja, aproximadamente singular, isto é, $\det(\mathbf{X}^T\mathbf{X}) \cong 0$, precisaremos de um método alternativo para solucionarmos (1.2) a fim de que os estimadores dos parâmetros não sejam inflacionados. O objetivo deste trabalho é apresentar um dos métodos alternativos, chamado regressão "ridge", o qual será visto no terceiro capítulo.

1.1.2 Propriedades

Devemos assumir algumas hipóteses para podermos analisar o modelo (1.1) estatisticamente. Essas hipóteses são chamadas condições de Gauss Markov (G-M)

$$E(\varepsilon_i) = 0$$

$$E(\varepsilon_i^2) = \sigma^2$$

$$E(\varepsilon_i\varepsilon_j) = 0 \quad i \neq j.$$

Na forma matricial teremos

$E(\varepsilon) = \mathbf{0}$ e $E(\varepsilon^T\varepsilon) = \sigma^2\mathbf{I}$, onde $\mathbf{0}$ representará o vetor de zeros.

As implicações imediatas destas hipóteses são:

i. $E(\mathbf{y}) = E(\mathbf{X}\beta + \varepsilon) = E(\mathbf{X}\beta) + E(\varepsilon) = \mathbf{X}\beta$

ii. $Cov(\mathbf{y}) = E(\mathbf{X}\beta + \varepsilon)\,(\mathbf{X}\beta + \varepsilon)^T = E(\varepsilon^T\varepsilon) = \sigma^2\mathbf{I}$

iii. $E(\mathbf{b}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^TE(\mathbf{y}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta = \beta$, se $\det(X^TX) \neq 0$.

Definição 3: Uma função $h(\mathbf{y})$ é não viciada para $f(\beta)$, se $E(h(\mathbf{y})) = f(\beta)$.

Definição 4: Uma função $f(\beta)$ é estimável, se existe $h(\mathbf{y})$ tal que $h(\mathbf{y})$ seja não viciado para $f(\beta)$, isto é, $E(h(\mathbf{y})) = f(\beta)$ para qualquer $\beta \in R^p$.

Desta forma em **iii**, sob as condições de G-M, **b** é um estimador não viciado de β.

A variância de **b** pode ser obtida através da matriz de variância-covariância, que é dada por:

iv. $\text{cov}(\mathbf{b}) = \text{cov}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y})$,

considerando $A = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$

$$\text{cov}(\mathbf{b}) = \text{cov}(\mathbf{Ay}) = \mathbf{A}\text{cov}(\mathbf{y})\mathbf{A}^T = \mathbf{A}\sigma^2\mathbf{I}\mathbf{A}^T = \sigma^2\mathbf{A}\mathbf{I}\mathbf{A}^T = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

Se denotarmos $\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1}$, a variância de b_j é $C_{jj}\sigma^2$ e a covariância entre b_i e b_j será $C_{ij}\sigma^2$, onde C_{jj} : corresponde ao j-ésimo elemento da diagonal principal da matriz $(\mathbf{X}^T\mathbf{X})^{-1}$ e C_{ij} : corresponde ao i-ésimo elemento da j-ésima coluna da mesma matriz.

Uma última consideração é de que os erros devem ser normalmente distribuídos. Essa pressuposição será tomada como verdadeira no decorrer de todo trabalho.

O modelo de regressão ajustado será dado por:

$$\hat{y} = \mathbf{Xb}$$

logo, $\hat{y} = \mathbf{Xb} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{Hy}$, onde $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ é chamada matriz projeção.

A diferença entre o valor observado e o correspondente valor ajustado é o resíduo. Representado na forma matricial por:

$$\mathbf{e} = \mathbf{y} - \hat{y}.$$

Podemos também representar o resíduo pela expressão:

$$\mathbf{e} = \mathbf{y} - \mathbf{Hy} = (\mathbf{I} - \mathbf{H})\mathbf{y}, \text{ considere } \mathbf{M} = \mathbf{I} - \mathbf{H}.$$

Teorema 1: A matriz **H** e **M** são simétricas e idempotentes, isto é, satisfazem as seguintes propriedades $\mathbf{H}^T = \mathbf{H}$ e $\mathbf{HH} = \mathbf{H}$, respectivamente.

Dem.:

$$\mathbf{H}^T = (\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{H}$$

$$\mathbf{HH} = (\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{H}$$

$$\mathbf{M}^T = (\mathbf{I} - \mathbf{H})^T = \mathbf{I} - \mathbf{H} = \mathbf{M}$$

$$\mathbf{MM} = (\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) = \mathbf{I} - \mathbf{H} - \mathbf{H} + \mathbf{HH} = \mathbf{I} - \mathbf{H} - \mathbf{H} + \mathbf{H} = \mathbf{I} - \mathbf{H} = \mathbf{M} \quad \blacksquare$$

Teorema 2: (Teorema de Gauss Markov)

No modelo $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ sob as condições de G-M, o estimador linear não viciado de variância mínima da função estimável $\mathbf{l}^T\beta$ é $\mathbf{l}^T\mathbf{b}$, onde **b** é solução da equação normal (1.4) e **l** é vetor $(p \times 1)$.

Dem.:

Seja $\mathbf{q}^T\mathbf{y}$, onde **q** é vetor $(n \times 1)$, outro estimador linear não viciado de $\mathbf{l}^T\beta$, então,

$$\mathbf{l}^T\beta = E(\mathbf{q}^T\mathbf{y}) = \mathbf{q}^T E(\mathbf{y}) = \mathbf{q}^T\mathbf{X}\beta \ , \forall\ \beta$$

logo,

$$\mathbf{l}^T = \mathbf{q}^T\mathbf{X}$$

$$\text{var}(\mathbf{q}^T\mathbf{y}) = \mathbf{q}^T\text{cov}(\mathbf{y})\mathbf{q} = \mathbf{q}^T(\sigma^2\mathbf{I})\mathbf{q} = \sigma^2\mathbf{q}^T\mathbf{q}$$

$$\text{var}(\mathbf{l}^T\mathbf{b}) = \mathbf{l}^T\text{cov}(\mathbf{b})\mathbf{l} = \sigma^2\mathbf{l}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{l} = \sigma^2\mathbf{q}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{q} \text{ e}$$

$$\text{var}(\mathbf{q}^T\mathbf{y}) - \text{var}(\mathbf{l}^T\mathbf{b}) = \sigma^2\mathbf{q}^T\mathbf{q} - \sigma^2\mathbf{q}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{q} =$$

$$= \sigma^2\mathbf{q}^T(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{q} =$$

$$= \sigma^2\mathbf{q}^T(\mathbf{I} - \mathbf{H})\mathbf{q} = \sigma^2\mathbf{q}^T(\mathbf{I} - \mathbf{H})^T(\mathbf{I} - \mathbf{H})\mathbf{q} = \sigma^2 \left\|(\mathbf{I} - \mathbf{H})\mathbf{q}\right\|^2 \geq 0$$

Portanto,

$$\text{var}(\mathbf{q}^T\mathbf{y}) \geq \text{var}(\mathbf{l}^T\mathbf{b}) \qquad \blacksquare$$

Logo, qualquer combinação linear do estimador de mínimos quadrados, $\mathbf{l}^T\mathbf{b}$, são estimadores lineares não viciados de variância mínima de $\mathbf{l}^T\beta$. Este é um importante resultado, pois, garante que estimadores de mínimos quadrados são de menor variância dentre os estimadores lineares não viciados.

1.1.3 Estimação de σ^2

Considerando-se $\mathbf{e}^T\mathbf{e}$ a soma de quadrados do resíduo, denotada por SQE como um estimador de σ^2

$$SQE = \mathbf{e}^T\mathbf{e} = \mathbf{y}^T(\mathbf{I} - \mathbf{H})^T(\mathbf{I} - \mathbf{H})\mathbf{y}$$

valendo-se do fato de $\mathbf{I} - \mathbf{H}$ ser idempotente e simétrica, teremos:

$$SQE = \mathbf{y}^T(\mathbf{I} - \mathbf{H})\,\mathbf{y}$$

$$E(\mathbf{e}^T\mathbf{e}) = E(SQE) = E[\mathbf{y}^T(\mathbf{I} - \mathbf{H})\,\mathbf{y}].$$

Usando o resultado de que o valor esperado de uma forma quadrática que é dado por:

$$E(\mathbf{y}^T\mathbf{A}\mathbf{y}) = \text{tr}(\mathbf{A}\Sigma) + \mu^T\mathbf{A}\mu, \text{ onde } E(\mathbf{y}) = \mu \text{ e Var } (\mathbf{y}) = \Sigma \qquad (1.3)$$

teremos que:

$$E[\mathbf{y}^T(\mathbf{I} - \mathbf{H})\,\mathbf{y}] = \sigma^2\,\text{tr}(\mathbf{I} - \mathbf{H}) + \beta^T\mathbf{X}^T(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{X}\beta =$$

$$= \sigma^2 [tr(\mathbf{I}) - tr(\mathbf{H})] + \beta^T \mathbf{X}^T \mathbf{X} \beta - \beta^T X^T X (X^T X)^{-1} X^T X \beta =$$

$$= \sigma^2 (n - p - 1) + \beta^T \mathbf{X}^T \mathbf{X} \beta - \beta^T \mathbf{X}^T \mathbf{X} \beta =$$

$$= \sigma^2 (n - p - 1)$$

logo,

$$E(SQE) = (n - p - 1)\sigma^2$$

então, defini-se o estimador da variância do erro, como:

$$\hat{\sigma}^2 = \frac{SQE}{n - p - 1}$$

Esta relação também é denominada quadrado médio do erro (QME), pois qualquer soma de quadrado dividido por seu respectivo grau de liberdade é chamada quadrado médio. Os graus de liberdade da soma de quadrado do erro são n – p – 1 que corresponde ao número da amostra menos o número de parâmetros no modelo.

Podemos observar que o estimador acima independe do modelo considerado.

1.1.4 Análise da Variância

A soma de quadrados são obtidas da seguinte maneira, considerando a relação:

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

elevando ao quadrado ambos os membros, teremos:

$$(y_i - \bar{y})^2 = (\hat{y}_i - \bar{y})^2 + (y_i - \hat{y}_i)^2 + 2(\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$$

e somando para i de 1 até n:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n} (\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$$

como,

$$2\sum_{i=1}^{n} (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = 0$$

então,

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

O primeiro termo à esquerda corresponde a soma de quadrados total, ou seja:

$$SQT = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = y^T y - n\bar{y}^2,$$

primeiro termo à direita é a soma de quadrados da regressão

$$SQR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} \hat{y}_i^2 - n\bar{y}^2 = \hat{y}^T \hat{y} - n\bar{y}^2$$

e o segundo termo à direita é a soma de quadrados do erro

$$SQE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \Sigma e_i^2 = \mathbf{e}^T\mathbf{e}.$$

Desta forma, teremos que:

$$SQT = SQR + SQE \tag{1.4}$$

isto indica que a soma de quadrados possui a propriedade de adição.

Neste caso, os graus de liberdade da regressão são iguais a p que relaciona-se ao número de regressoras. Os graus de liberdade também possuem a propriedade de aditividade, assim, os graus de liberdade da soma de quadrados total são n – 1.

O valor esperado da soma de quadrados da regressão é o seguinte:

$$SQR = y^T\mathbf{H}y - (1/n)\mathbf{y}^T\mathbf{J}\mathbf{J}^T\mathbf{y} = \mathbf{y}^T(\mathbf{H} - \frac{\mathbf{J}\mathbf{J}^T}{n})\mathbf{y}, \text{ onde } \mathbf{J} = (1\ 1 \ldots 1)^T.$$

utilizando o valor esperado de uma forma quadrática, teremos:

$$E(SQR) = E[y^T(\mathbf{H} - \frac{\mathbf{J}\mathbf{J}^T}{n})y] = tr(\mathbf{H} - \frac{\mathbf{J}\mathbf{J}^T}{n})\sigma^2 + \beta^T\mathbf{X}^T(\mathbf{H} - \frac{\mathbf{J}\mathbf{J}^T}{n})\mathbf{X}\beta$$

$$= \sigma^2[p + 1 - \frac{n}{n}] + \beta^T\mathbf{X}^T(\mathbf{H} - \frac{\mathbf{J}\mathbf{J}^T}{n})\mathbf{X}\beta$$

$$= \sigma^2 p + \beta^T\mathbf{X}^T(\mathbf{H} - \frac{\mathbf{J}\mathbf{J}^T}{n})\mathbf{X}\beta. \tag{1.5}$$

Agora o valor esperado da soma de quadrados total, será a esperança de:

$$SQT = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\frac{\mathbf{J}\mathbf{J}^T}{n}\mathbf{y} = \mathbf{y}^T(I - \frac{\mathbf{J}\mathbf{J}^T}{n})\mathbf{y}$$

logo,

$$E(SQT) = E[\mathbf{y}^T(I - \frac{JJ^T}{n})\mathbf{y}].$$

Novamente utilizando o valor esperado de uma forma quadrática, como em (1.3), teremos:

$$E[\mathbf{y}^T(\mathbf{I} - \frac{JJ^T}{n})\mathbf{y}] = tr[\sigma^2(\mathbf{I} - \frac{JJ^T}{n})] + \beta^T\mathbf{X}^T(I - \frac{JJ^T}{n})\mathbf{X}\beta$$

$$= (n-1)\sigma^2 + \beta^T\mathbf{X}^T(\mathbf{I} - \frac{JJ^T}{n})\mathbf{X}\beta \qquad (1.6)$$

Observemos agora que se dividirmos os valores esperados da soma de quadrados da regressão e da soma de quadrados total pelo seus respectivos graus de liberdade obteremos os valores esperados dos quadrados médios. Sendo assim, as expressões (1.5) e (1.6) mostram que estas estimativas referidas seriam estimadores viciados da variância, enquanto que QME será sempre um estimador não viciado para σ^2.

Podemos resumir esta análise em uma tabela de análise da variância.

Tabela 1: Análise da variância para o modelo corrigido pela média

Fonte	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	Esperança do Quadrado Médio
Regressão	p	$SQR = \mathbf{y}^T\mathbf{H}\mathbf{y} - n\bar{y}^2$	QMR=SQR/p	$\sigma^2 + \frac{1}{p}\beta^T\mathbf{X}^T(\mathbf{H} - \frac{JJ^T}{n})\mathbf{X}\beta$
Erro	n−p−1	$SQE = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}\mathbf{y}$	QME=SQE/(n-p-1)	σ^2
Total	n−1	$\mathbf{y}^T\mathbf{y} - n\bar{y}^2$		

1.2 Coeficiente de Determinação

Se isolarmos a SQR em (1.4) e dividirmos por SQT ambos os lados, teremos

$$\frac{SQR}{SQT} = 1 - \frac{SQE}{SQT} \tag{1.7}$$

O primeiro membro de (1.7) é a proporção de variabilidade de y explicada pelo modelo de regressão. O lado direito consistirá: um menos a variabilidade não explicada. Este conceito terá um nome especial, definiremos R^2, o coeficiente de determinação, por

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT}. \tag{1.8}$$

Como $0 \leq SQE \leq SQT$ segue que R^2 assume valores entre 0 e 1. Quando os valores de R^2 são próximos de 1 implicam que a variabilidade de y é altamente explicada pelo modelo de regressão. Contudo, há casos em que R^2 grande é consequência da adição de um termo ao modelo, e necessariamente, não significa que o novo modelo é melhor explicado comparado ao primeiro. Por isso, observamos que não devemos tirar conclusões baseado somente no valor de R^2.

A magnitude de R^2 depende do campo de variação de **X**. Geralmente R^2 aumenta com o crescimento da dispersão em **X**, e diminui com o decréscimo da dispersão em **X**. Esta afirmação é consequência imediata da equação (1.8), pois, R^2 está diretamente relacionado com a soma de quadrados da regressão, que corresponde a dispersão de **X**.

Segundo *Montgomery* (1992), Hahn observou que o valor esperado de R^2 é aproximadamente:

$$E(R^2) \cong \frac{b_1^2 SQR}{b_1^2 SQR + \sigma^2}$$

Nesta relação vemos que o valor esperado de R^2 crescerá quando SQR (uma medida de extensão dos **x**'s) aumenta, analogamente R^2 diminui quando a SQR decresce. O que confirma que a magnitude de R^2 depende do campo de variação de **X**.

É importante notar que R^2 não mede aproximação do modelo linear, isto é, podemos ter R^2, razoavelmente, grande e isto não significa que exista relação linear; pode existir, por exemplo, uma relação cúbica entre as variáveis **y** e **x**.

Se considerarmos o modelo (1.1) com p=1, obteremos um modelo de regressão linear simples, isto é,

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i \quad i=1,2,...,n$$

definiremos o coeficiente de correlação, r_{xy}, como:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Por facilidade de expressão consideraremos $S_{xy}^2 = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$, $S_x^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$ e $S_y^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2$, então teremos:

$$r_{xy} = \frac{S_{xy}^2}{S_x S_y}.$$

O r_{xy} indica a relação linear existente entre as variáveis x e y; este coeficiente assume valores entre −1 e 1. Valores próximos de −1 ou 1 indicam forte relação linear entre as variáveis x e y. Vale ressaltar que r_{xy} próximo de zero indica que não existe relação linear entre as variáveis, entretanto, isto não quer dizer que não exista qualquer outra relação entre elas. Novamente, pode haver uma relação quadrática ou cúbica.

Notamos, neste caso de regressão linear simples que o coeficiente de determinação é igual ao coeficiente de correlação.

$$R^2 = \frac{SQR}{SQT} = \frac{S_{xy}^2}{S_{xx}SQT} = \frac{S_{xy}^2}{S_{xx}S_{yy}} = r_{xy}^2$$

1.3 Centralização e Escalonamento

No decorrer deste trabalho trataremos da matriz **X** centrada e escalonada. Dizemos que a matriz **X** é centrada e escalonada se de cada elemento da matriz é subtraído pela média da coluna e dividido pela raiz quadrada da soma de quadrados dos desvios com relação a média, ou seja, dividido por $S_j = \sqrt{\sum_i (x_i - \bar{x}_j)^2}$.

A motivação para tal procedimento se deve a:

1) Redução do erro de arredondamento na inversão da matriz X^TX.

2) Possível aumento da explicabilidade das variáveis e seus coeficientes de regressão.

3) A possibilidade de compararmos diretamente os coeficientes de regressão das diferentes variáveis.

Por exemplo[1]: Suponhamos o ajustado $\hat{y} = -171 + 1.92\mathbf{x_1} + 0.286\mathbf{x_2}$ com

y: capacidade pulmonar em centilitros

$\mathbf{x_1}$: altura em centímetro

$\mathbf{x_2}$: peso em quilogramas

Não faz sentido comparar os coeficientes 1.92 com 0.286, pois estão em diferentes escalas de medição. Agora, se padronizarmos as variáveis regressoras obteremos a seguinte equação estimada $\hat{y} = 193 + 12.9\mathbf{w_1} + 3.28\mathbf{w_2}$. Dessa maneira podemos comparar 12.9 e 3.28 para concluir que a diferença nas capacidades pulmonares são mais influenciadas pelas alturas do que pelos pesos.

[1] Exemplo retirado do livro *Birkes e Dodge* (1993), pág. 177-178.

Com isso, teremos a padronização da forma:

$$X=\begin{pmatrix} 1 & \frac{x_{11}-\bar{x}_1}{S_1} & \cdots & \frac{x_{1p}-\bar{x}_p}{S_p} \\ 1 & \frac{x_{21}-\bar{x}_1}{S_1} & \cdots & \frac{x_{2p}-\bar{x}_p}{S_p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \frac{x_{n1}-\bar{x}_1}{S_1} & \cdots & \frac{x_{np}-\bar{x}_p}{S_p} \end{pmatrix}.$$

Observe que não centramos e nem escalonamos a coluna que corresponde ao intercepto, pois, se assim fosse teríamos uma coluna de zeros. Deste modo, a matriz **X** pode ser representada por:

$$\mathbf{X} = (\mathbf{J}\ \mathbf{W})$$

onde, **J** é a coluna de uns e **W** é matriz centrada e escalonada sem a coluna de uns.

Pode-se provar que $\mathbf{W}^T\mathbf{W}$ terá forma da matriz correlação.

$$\mathbf{W}^T\mathbf{W} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix},$$

r_{ij} corresponde a correlação entre o elemento da i-ésima linha e a j-ésima coluna.

De fato:

Suponha p = 2. Neste caso $\mathbf{W} = \begin{pmatrix} \frac{x_{11}-\bar{x}_1}{\sqrt{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}} & \frac{x_{12}-\bar{x}_2}{\sqrt{\sum_{i=1}^{n}(x_{i1}-\bar{x}_2)^2}} \\ \vdots & \vdots \\ \frac{x_{n1}-\bar{x}_1}{\sqrt{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}} & \frac{x_{n2}-\bar{x}_2}{\sqrt{\sum_{i=1}^{n}(x_{i1}-\bar{x}_2)^2}} \end{pmatrix} = (\mathbf{w_1}\ \mathbf{w_2})$

então,

$$\mathbf{W^TW} = \begin{pmatrix} w_1^T w_1 & w_1^T w_2 \\ w_2^T w_1 & w_2^T w_2 \end{pmatrix}.$$

$$\mathbf{w_1}^T\mathbf{w_1} = \begin{pmatrix} \frac{x_{11}-\bar{x}_1}{\sqrt{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}} & \cdots & \frac{x_{n1}-\bar{x}_1}{\sqrt{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}} \end{pmatrix} \begin{pmatrix} \frac{x_{11}-\bar{x}_1}{\sqrt{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}} \\ \vdots \\ \frac{x_{n1}-\bar{x}_1}{\sqrt{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}} \end{pmatrix} =$$

$$= \frac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}{(\sqrt{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2})^2} = \frac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2} = 1 \qquad (1.9)$$

analogamente, teremos $\mathbf{w_2}^T\mathbf{w_2} = 1$.

Agora, $\mathbf{w_1}^T\mathbf{w_2} = \begin{pmatrix} \frac{x_{11}-\bar{x}_1}{\sqrt{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}} & \cdots & \frac{x_{n1}-\bar{x}_1}{\sqrt{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}} \end{pmatrix} \begin{pmatrix} \frac{x_{12}-\bar{x}_1}{\sqrt{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}} \\ \vdots \\ \frac{x_{n2}-\bar{x}_1}{\sqrt{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}} \end{pmatrix} =$

$$= \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{11} - \bar{x}_1)^2}\sqrt{\sum_{i=1}^{n}(x_{i2} - \bar{x}_2)^2}} = r_{12}$$

analogamente, teremos $w_2^T w_1 = r_{21}$

Assim:

$$\mathbf{W^T W} = \begin{pmatrix} 1 & r_{12} \\ r_{21} & 1 \end{pmatrix}.$$

Suponhamos que seja verdade para p = 1, 2 , ... , k e provemos que vale para p = k+1.

Logo, teremos que $\mathbf{W^T W} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{21} & 1 & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$ (1.10)

devemos provar que $\mathbf{W^T W} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1k} & r_{1,k+1} \\ r_{21} & 1 & \cdots & r_{2k} & r_{2,k+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ r_{k1} & r_{k2} & \cdots & 1 & r_{k,k+1} \\ r_{k+1,1} & r_{k+1,2} & \cdots & r_{k+1,k} & 1 \end{pmatrix}$.

Para isso considere **W** particionado da forma: **W** = (**M** **m**), onde **m** corresponde a última coluna de **W** e **M** as k primeiras colunas. Então:

$$\mathbf{W^T W} = \begin{pmatrix} \mathbf{M^T M} & \mathbf{M^T m} \\ \mathbf{m^T M} & \mathbf{m^t m} \end{pmatrix}$$

$\mathbf{M}^T\mathbf{M}$ por hipótese de indução é igual a (1.11) e $\mathbf{m}^T\mathbf{m} = 1$ como em (1.9), devemos verificar o vetor $\mathbf{M}^T\mathbf{m}$.

$$\mathbf{M} = \begin{pmatrix} \dfrac{x_{11} - \bar{x}_1}{\sqrt{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}} & \dfrac{x_{12} - \bar{x}_2}{\sqrt{\sum_{i=1}^{n}(x_{i2} - \bar{x}_2)^2}} & \cdots & \dfrac{x_{1k} - \bar{x}_k}{\sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}} \\ \vdots & \vdots & \vdots & \vdots \\ \dfrac{x_{n1} - \bar{x}_1}{\sqrt{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}} & \dfrac{x_{n2} - \bar{x}_2}{\sqrt{\sum_{i=1}^{n}(x_{i2} - \bar{x}_2)^2}} & \cdots & \dfrac{x_{nk} - \bar{x}_k}{\sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}} \end{pmatrix}$$

e

$$\mathbf{m} = \begin{pmatrix} \dfrac{x_{1,k+1} - \bar{x}_{k+1}}{\sqrt{\sum_{i=1}^{n}(x_{ik+1} - \bar{x}_{k+1})^2}} \\ \vdots \\ \dfrac{x_{nk} - \bar{x}_{k+1}}{\sqrt{\sum_{i=1}^{n}(x_{ik+1} - \bar{x}_{k+1})^2}} \end{pmatrix}.$$

$$\text{Logo, } \mathbf{M}^T\mathbf{m} = \begin{pmatrix} \dfrac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i,k+1} - \bar{x}_{k+1})}{\sqrt{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}\sqrt{\sum_{i=1}^{n}(x_{i,k+1} - \bar{x}_{k+1})^2}} \\ \vdots \\ \dfrac{\sum_{i=1}^{n}(x_{ik} - \bar{x}_1)(x_{i,k+1} - \bar{x}_{k+1})}{\sqrt{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}\sqrt{\sum_{i=1}^{n}(x_{i,k+1} - \bar{x}_{k+1})^2}} \end{pmatrix} = \begin{pmatrix} r_{1,k+1} \\ \vdots \\ r_{k,k+1} \end{pmatrix}$$

e analogamente teremos $\mathbf{m}^T\mathbf{M} = \begin{pmatrix} r_{k+1,1} \\ \vdots \\ r_{k+1,k} \end{pmatrix}^T$.

Portanto fica provado que $\mathbf{W}^T\mathbf{W}$ tem forma de uma matriz correlação quando W é uma matriz padronizada. ■

1.4 Decomposição de Valores Singulares

Uma matriz $\mathbf{X}$ (n×p) pode ser decomposta na forma $\mathbf{X} = \mathbf{UDV}^T$, onde $\mathbf{U}$ matriz (n×p), $\mathbf{D}$ matriz (p×p) e $\mathbf{V}$ matriz (p×p), tal que $\mathbf{U}^T\mathbf{U} = \mathbf{I}_n$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}_p$ e $\mathbf{D} = \text{diag}(\sqrt{\lambda_1},\sqrt{\lambda_2},...,\sqrt{\lambda_p})$ é a matriz diagonal cujos elementos são chamados valores singulares.

Estamos interessados na decomposição da matriz $\mathbf{X}^T\mathbf{X}$, então

$$\mathbf{X}^T\mathbf{X} = (\mathbf{UDV}^T)^T(\mathbf{UDV}^T) = \mathbf{VDU}^T\mathbf{UDV}^T = \mathbf{VDI}_p\mathbf{DV}^T = \mathbf{VD}^2\mathbf{V}^T$$

onde, $\mathbf{V}$ é tal que $\mathbf{V}^T\mathbf{V} = \mathbf{I}_p$, $\mathbf{D}^2$ é a matriz cuja diagonal são os quadrados dos valores singulares que são os autovalores da matriz $\mathbf{X}^T\mathbf{X}$. As colunas de V são os autovetores de $\mathbf{X}^T\mathbf{X}$ associados com os p autovalores.

Podemos rescrever da forma $\mathbf{X}^T\mathbf{X} = \mathbf{V\Lambda V}^T$, onde $\mathbf{V}$ matriz de autovetores e $\mathbf{D}^2=\Lambda=\text{diag}(\lambda_1,\lambda_2,...,\lambda_p)$ são os autovalores da matriz $\mathbf{X}^T\mathbf{X}$.

Através da teoria de raiz característica, ver *Rao* (1973), teremos

$$\mathbf{X}^T\mathbf{X}\mathbf{v}_i = \lambda_i \mathbf{v}_i$$

$$\lambda_i = \mathbf{v}_i^T\mathbf{X}^T\mathbf{X}\mathbf{v}_i$$

$$\lambda_i = (\mathbf{X}\mathbf{v}_i)^T(\mathbf{X}\mathbf{v}_i)$$

$$\lambda_i = \| \mathbf{X}\mathbf{v}_i \|^2 \geq 0.$$

Logo , os autovalores da matriz $\mathbf{X}^T\mathbf{X}$ são não negativos.

Capítulo 2

Multicolinearidade

Em muitas análises de modelos de regressão deparamo-nos com o problema de mal condicionamento da matriz de delineamento. O efeito deste mal condicionamento é a inflação da variância do estimador de mínimos quadrados dos parâmetros e, possivelmente, dos valores preditos, ocorrendo também uma restrição na generalidade e aplicabilidade do modelo estimado.

Não é fácil identificar com precisão o efeito, separadamente, das variáveis envolvidas na correlação. Por isso, no intuito de minimizar o problema de uma maneira simples, eliminam-se variáveis do modelo que são menos significantes. Essa eliminação, muitas vezes, faz com que seja grande a perda de informação.

Uma vez detectado o mal condicionamento uma boa solução seria obter e incorporar mais informações ao modelo. Estas informações adicionais podem ser refletidas sob a forma de novos dados. Infelizmente, a possibilidade de resolvermos o problema por este procedimento é muito limitado.

Para o pesquisador incapaz de obter mais informações, alguns procedimentos tem sido desenvolvidos como, por exemplo, a regressão "ridge". Estes métodos nos proporcionam mais informações da amostra e produzem estimadores mais precisos.

Neste capítulo, discutiremos a dimensão do problema de multicolinearidade e procedimentos que podem ser usados para sua identificação.

2.1 O que é Multicolinearidade?

Definiremos multicolinearidade exata em função da dependência linear das colunas de **X**. Sendo $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_p$ as colunas da matriz **X**, podemos dizer que **X** está com o problema de multicolinearidade exata se as colunas de **X** são linearmente dependentes, isto é, se existe conjunto de constantes $t_1,t_2,\ldots,t_p$ não todas nulas, tal que:

$$\sum_{i=1}^{p} t_i \mathbf{x}_i = 0 \qquad (2.1)$$

Neste caso uma das variáveis pode ser determinada pelas outras e $\mathbf{X}^T\mathbf{X}$ será singular. Na prática tais situações são raras; agora, uma situação mais comum é quando a matriz possui multicolinearidade aproximada, ou seja, as colunas de **X** estão próximas da dependência linear e $\lambda_i > 0$, $\forall$ i = 1,2,...,p.

$$\sum_{i=1}^{p} t_i \mathbf{x}_i \approx 0 \qquad (2.2)$$

onde $\approx$ denota a proximidade. Portanto, teremos que uma das variáveis, digamos $\mathbf{x}_p$, pode ser, aproximadamente, determinada pelas outras.

$$\mathbf{x}_p \approx -\frac{\sum_{i \neq p} t_i \mathbf{x}_i}{t_p} \qquad (2.3)$$

Na situação descrita por (2.3) dizemos que **X** está mal condicionada e a matriz $\mathbf{X}^T\mathbf{X}$ será aproximadamente singular.

Um diagnóstico simples é o coeficiente correlação múltipla, este número é calculado da regressão do $\mathbf{x}_p$ nos outros $\mathbf{x}$'s. Se $R_p^2 = 1$ dizemos que $\mathbf{X}$ tem multicolinearidade exata, se $R_p^2 = 0$ então $\mathbf{X}$ é ortogonal, e quando R_p^2 é próximo de um dizemos que a matriz é aproximadamente multicolinear.

2.2 Efeitos da Multicolinearidade

Na seção anterior definiu-se multicolinearidade aproximada em termos da dependência linear entre as colunas de $\mathbf{X}$, segundo *Wetherill* (1986), podemos redefinir em termos da existência de um vetor unitário $\mathbf{t}$ (isto é, $\mathbf{t}^T\mathbf{t} = 1$) tal que:

$$\sum_{i=1}^{p} t_i \mathbf{x}_i = \delta$$

onde, δ é pequeno, isto é, $\| \delta \|^2 = \delta^T\delta < \varepsilon^2$, para ε suficientemente pequeno. Por fim,

$$\varepsilon > \|\delta\| = \| \sum_{i=1}^{p} t_i \mathbf{x}_i \| = (\mathbf{t}^T\mathbf{X}^T\mathbf{X}\mathbf{t})^{1/2}$$

onde, $\mathbf{t}^T = (t_1 , t_2 , \ldots , t_p)$

$$\mathbf{t}^T\mathbf{X}^T\mathbf{X}\mathbf{t} = \lambda < \varepsilon^2$$

De fato, podemos considerar $\mathbf{t} = \mathbf{V}\gamma$ para algum vetor γ ($p \times 1$) e $\mathbf{V}$ como na seção (1.4), empregando a decomposição de valores singulares, teremos:

$$\mathbf{t}^T\mathbf{X}^T\mathbf{X}\mathbf{t} = \gamma^T\mathbf{V}^T\mathbf{X}^T\mathbf{X}\mathbf{V}\gamma = \gamma^T\Lambda\gamma = \sum_{i=1}^{p}\gamma_i^2\lambda_i = \lambda, \text{ onde } \lambda_i \text{ é autovalor de } \mathbf{X}^T\mathbf{X}.$$

Agora,

$$\lambda = \sum_{i=1}^{p}\gamma_i^2\lambda_i \geq \min_{1\leq i\leq p}(\lambda_i)\sum_i\gamma_i^2 = \min_{1\leq i\leq p}(\lambda_i)$$

desde que

$$\sum_{i=1}^{p}\gamma_i^2 = \gamma^T\gamma = \mathbf{t}^T\mathbf{V}^T\mathbf{V}\mathbf{t} = \mathbf{t}^T\mathbf{t} = 1 \text{ e } \lambda_i > 0, \ \forall\ i=1,2,...,p,$$

então,

$$\min_{1\leq i\leq p}(\lambda_i) \leq \lambda < \varepsilon^2, \text{ para algum } \varepsilon \text{ pequeno}$$

Concluímos assim, que na presença de multicolinearidade aproximada há pelo menos um autovalor próximo de zero. Da decomposição de autovalores de $\mathbf{X}^T\mathbf{X}$, teremos:

$$(\mathbf{X}^T\mathbf{X})^{-1} = \mathbf{V}\Lambda^{-1}\mathbf{V}^T = \sum_{i=1}^{p}B_i \text{ , onde } B_i \text{ é a matriz definida por } \lambda_i^{-1}\mathbf{v}_i\mathbf{v}_i^T. \qquad (2.4)$$

O estimador de mínimos quadrados pode ser escrito da seguinte forma:

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \sum_{i=1}^{p}\lambda_i^{-1}\mathbf{v}_i d_i, \text{ onde } d_i = \mathbf{v}_i^T\mathbf{X}^T\mathbf{y} \qquad (2.5)$$

Por conveniência, assumimos que $\lambda_{min} = \lambda_p$, tal que $\lambda_1 > \lambda_2 > \ldots > \lambda_p$ então de (2.5) **b** será dominado por $\mathbf{v}_p$. Dado que $\mathbf{v}_p$ é autovetor de $\mathbf{X}^T\mathbf{X}$ temos que $\mathbf{X}^T\mathbf{X}\mathbf{v}_p = \lambda_p\mathbf{v}_p$ de modo que

$$\mathbf{v}_p^T\mathbf{X}^T\mathbf{X}\mathbf{v}_p = \lambda_p\mathbf{v}_p^T\mathbf{v}_p = \lambda_p$$

$$\Rightarrow \|\mathbf{X}\mathbf{v}_p\|^2 = (\mathbf{X}\mathbf{v}_p)^T(\mathbf{X}\mathbf{v}_p) = \lambda_p \leq \lambda < \varepsilon^2 \text{, para algum } \varepsilon \text{ pequeno}$$

Por conseguinte, a norma de $\mathbf{X}\mathbf{v}_p$ é tão pequena quanto o autovalor correspondente ao autovetor. Entretanto, se há mais que uma multicolinearidade aproximada existirá também mais que um autovalor pequeno e as variáveis envolvidas poderão ser identificadas usando os autovetores apropriados, como foi feito acima.

Considere agora a matriz de variância e covariância de **b**.

$$\text{Var}(\mathbf{b}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

usando (2.3), teremos:

$$\text{Var}(\mathbf{b}) = \sigma^2 \sum_{i=1}^{p} \lambda_i^{-1}\mathbf{v}_i\mathbf{v}_i^T$$

Observamos que a matriz de variância-covariância do estimador de mínimos quadrados é inflacionada com o mal condicionamento da matriz **X**, pois com o mal condicionamento demostramos a existência de algum autovalor próximo de zero, o que faz com que a variância do estimador, que é a diagonal de Var(**b**), seja muito grande. O mesmo ocorre com o erro quadrático médio total do estimador.

$$\text{EQMT}(\mathbf{b}) = \text{E}[(\mathbf{b} - \beta)^T(\mathbf{b} - \beta)]$$

$$= \text{E}[\text{tr}(\mathbf{b} - \beta)^T(\mathbf{b} - \beta)]$$

$$= \text{tr}[\text{Var}(\mathbf{b})]$$

$$= \sigma^2 tr[\mathbf{V}\Lambda^{-1}\mathbf{V}^T]$$

então, teremos o EQMT igual a:

$$\sigma^2 \sum_{i=1}^{p} \lambda_i^{-1} .$$

Desta forma, todo estimador **b** de mínimos quadrados , apesar de ser um estimador linear não viciado com variância mínima, possui uma variância muito grande na presença de multicolinearidade aproximada, o que o torna um estimador não conveniente.

2.3 Medidas de Multicolinearidade

Vimos que quando existe mal condicionamento na matriz **X** os autovalores da matriz $\mathbf{X}^T\mathbf{X}$ se aproximam de zero e, consequentemente, as variâncias e o EQMT tornam-se muito inflacionadas, o que não é conveniente estatisticamente, tendo desta forma a necessidade de se minimizar a variância e o EQMT. Veremos nesta seção alguns dos métodos de detecção da multicolinearidade.

Considere a matriz **X** padronizada e particionada da forma $\mathbf{X} = [\, \mathbf{X}^* \; \mathbf{x}_p \,]$, onde $\mathbf{x}_p$ corresponde a última coluna de **X** e $\mathbf{X}^*$ as p − 1 primeiras colunas, e supomos X^* de posto completo.

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} (\mathbf{X}^*)^T \\ \mathbf{x}_p^T \end{bmatrix} [\, \mathbf{X}^* \; \mathbf{x}_p \,] = \begin{bmatrix} \mathbf{X}^{*T}\mathbf{X}^* & \mathbf{X}^{*T}\mathbf{x}_p \\ \mathbf{x}_p^T\mathbf{X}^* & \mathbf{x}_p^T\mathbf{x}_p \end{bmatrix}$$

A inversa de $\mathbf{X}^T\mathbf{X}$ é tal que $\mathbf{X}^T\mathbf{X}\,(\mathbf{X}^T\mathbf{X})^{-1} = \mathbf{I}$.

Suponhamos que $(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & c_p \end{bmatrix}$,

onde, **A** matriz $(p \times (p-1))$, **B** matriz $((p-1)\times 1)$ e c_p é uma constante, portanto:

$$\begin{bmatrix} \mathbf{X}^{*T}\mathbf{X}^* & \mathbf{X}^{*T}\mathbf{x}_p \\ \mathbf{x}_p^T\mathbf{X}^* & \mathbf{x}_p^T\mathbf{x}_p \end{bmatrix} \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & c_p \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0}^T & 1 \end{bmatrix},$$

ou seja,

$$\mathbf{X}^{*T}\mathbf{X}^*\mathbf{A} + \mathbf{X}^{*T}\mathbf{x}_p\mathbf{B}^T = \mathbf{I} \qquad (2.6)$$

$$\mathbf{X}^{*T}\mathbf{X}^*\mathbf{B} + \mathbf{X}^{*T}\mathbf{x}_p c_p = \mathbf{0} \qquad (2.7)$$

$$\mathbf{x}_p^T\mathbf{X}^*\mathbf{A} + \mathbf{x}_p^T\mathbf{x}_p\mathbf{B}^T = \mathbf{0}^T \qquad (2.8)$$

$$\mathbf{x}_p^T\mathbf{X}^*\mathbf{B} + \mathbf{x}_p^T\mathbf{x}_p c_p = 1 \qquad (2.9)$$

de (2.7) teremos :

$$\mathbf{X}^{*T}\mathbf{X}^*\mathbf{B} = -\mathbf{X}^{*T}\mathbf{x}_p c_p$$

pre-multiplicando ambos lados por $(\mathbf{X}^{*T}\mathbf{X}^*)^{-1}$,

$$\mathbf{B} = -(\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}\mathbf{x}_p c_p \qquad (2.10)$$

substituindo (2.10) em (2.9) teremos:

$$\mathbf{x}_p^T \mathbf{X}^*(-(\mathbf{X}^{*T}\mathbf{X}^*)^{-1} \mathbf{X}^{*T}\mathbf{x}_p c_p) + \mathbf{x}_p^T \mathbf{x}_p c_p = 1$$

isolando o termo comum, c_p:

$$[-\mathbf{x}_p^T \mathbf{X}^*(\mathbf{X}^{*T}\mathbf{X}^*)^{-1} \mathbf{X}^{*T}\mathbf{x}_p) + \mathbf{x}_p^T \mathbf{x}_p]c_p = 1$$

$$c_p = [\mathbf{x}_p^T \mathbf{x}_p - \mathbf{x}_p^T \mathbf{X}^*(\mathbf{X}^{*T}\mathbf{X}^*)^{-1} \mathbf{X}^{*T}\mathbf{x}_p]^{-1}. \qquad (2.11)$$

Logo, o p-ésimo termo da diagonal principal da inversa de $\mathbf{X}^T\mathbf{X}$ é dado por c_p em (2.11), como $\mathbf{X}$ foi centrada e escalonada, então o primeiro termo da expressão (2.11) é igual a um, como em (1.3), enquanto o segundo termo é a soma de quadrados regressão de $\mathbf{x}_p$ em todas outras variáveis.

Consequentemente, como por definição R_p^2 é dado por:

$$R_p^2 = \frac{SQR_p}{\mathbf{x}_p^T \mathbf{x}_p} = SQR_p$$

então, teremos que:

$$c_p = (1 - R_p^2)^{-1} \qquad (2.12)$$

onde, R_p^2 é o coeficiente de determinação da regressão de $\mathbf{x}_p$ em todas as outras variáveis regressoras restantes.

Se a p-ésima variável está envolvida na multicolinearidade temos que em (2.12) R_p^2 se aproxima de um, assim o coeficiente determinação, R_p^2, pode ser usado para indicar quais são as variáveis envolvidas na multicolinearidade.

2.3.1 Fator de Inflação da Variância (VIF)[1]

Segundo *Berk* (1977), o termo fator de inflação da variância foi atribuído por Marquard em 1960. Ele estabeleceu o nome de VIF pelo crescimento da variância quando os dados são não-ortogonais comparando-os aos dados ortogonais.

O VIF de uma variável $\mathbf{x}_i$ mede o quanto esta se relaciona linearmente com as outras variáveis regressoras. Especificamente,

$$VIF_i = \frac{1}{1 - R_i^2}$$

onde, R_i^2 é o coeficiente de determinação da regressão de $\mathbf{x}_i$ nas outras variáveis.

De (2.11) e (2.12) vemos que quando a matriz $\mathbf{X}$ é centrada e escalonada o i-ésimo fator de inflação da variância é exatamente igual ao i-ésimo elemento da diagonal de $(\mathbf{X}^T\mathbf{X})^{-1}$. Neste caso, teremos a variância do estimador de mínimos quadrados igual a:

$$Var(b_i) = \sigma^2 VIF_i.$$

[1] VIF é do inglês *Variance Inflation Factor*

Há também uma outra relação entre o VIF e a variância do estimador de mínimos quadrados do coeficiente de regressão visto que podemos mostrar quando a matriz **X** não está padronizada, escrevendo:

$$\mathrm{Var}(b_i) = \frac{\sigma^2}{\frac{s_i^2}{n-1}} \mathrm{VIF}_i \qquad (2.13)$$

De fato, notamos que o último elemento da diagonal principal de $(\mathbf{X}^T\mathbf{X})^{-1}$ é $c_p = [\,\mathbf{x}_p^T\mathbf{x}_p - \mathbf{x}_p^T\mathbf{X}^*(\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}\mathbf{x}_p\,]^{-1}$. Podemos notar que essa quantidade é o inverso da soma de quadrados residual de uma análise de regressão de $\mathbf{x}_p$ como função linear de $\mathbf{x}_1, \ldots, \mathbf{x}_{p-1}$, isto é, se considerarmos o modelo $\mathbf{x}_p = \beta\mathbf{X}^* + \varepsilon$ temos que a soma de quadrado residual é $\mathbf{x}_p^T\mathbf{x}_p - \mathbf{x}_p^T\mathbf{H}\mathbf{x}_p$ neste caso $\mathbf{H} = \mathbf{x}_p^T\mathbf{X}^*(\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}\mathbf{x}_p$.

Logo:

$$\mathrm{Var}(b_p) = \sigma^2 c_p$$

porém, pela fórmula (2.11) teremos

$$\mathrm{Var}(b_p) = \frac{\sigma^2}{\mathrm{SQE}_p}.$$

Provaremos que

$$\frac{\sigma^2}{\mathrm{SQE}_p} = \frac{\sigma^2 \mathrm{VIF}_p}{(n-1)S_p^2}. \qquad (2.14)$$

Sabemos que:

$$VIF_i = \frac{1}{1 - R_i^2} \text{ e } S_p^2 = \frac{\sum_i (x_{ip} - \bar{x}_p)^2}{(n-1)}$$

substituindo em (2.14) teremos:

$$\frac{1}{SQE_p} = \frac{1}{(1 - R_p^2)} \Rightarrow R_p^2 = 1 - \frac{SQE_p}{\sum_i (x_{ip} - \bar{x}_p)^2},$$

como definido na seção 1.3.

Enfim, a igualdade (2.14) é verdadeira , comprovando a equação (2.13). ■

O VIF é uma medida importante para diagnosticar a multicolinearidade, pois, VIF alto indica que R_i^2 se aproxima de um e, consequentemente, aponta para colinearidade aproximada.

2.3.2 Índice de Condição

Os autovalores da matriz $\mathbf{X}^T\mathbf{X}$ também são importantes para o diagnóstico de multicolinearidade, pois sabemos que quando há colinearidade aproximada entre as regressoras existem autovalores próximos de zero.

Desta forma em 1948, Turing introduziu o que chamamos de número de condição de $\mathbf{X}^T\mathbf{X}$, que é definido por

$$\eta = \frac{\lambda_{max}}{\lambda_{min}}.$$

Geralmente se o número de condição é menor que 100, inexiste problema sério de colinearidade. Números de condição entre 100 e 1000 implica numa colinearidade moderada quase forte e se n excede 1000 implica numa colinearidade forte.

O índice de condição da matriz $\mathbf{X}^T\mathbf{X}$ é um conjunto de p valores da forma:

$$\eta_i = \frac{\lambda_{max}}{\lambda_i} \quad i=1,2,\ldots,p$$

onde, λ_i são os autovalores de $\mathbf{X}^T\mathbf{X}$.

O i-ésimo maior valor de η_i é um limite superior aproximado para o número de condição da matriz de correlação formado pela eliminação da coluna i de $\mathbf{X}$. Assim, existem tantas multicolinearidades em $\mathbf{X}$ quanto valores grandes de η_i.

2.4 Solução para Multicolinearidade

Detectada a presença de mal condicionamento, uma alternativa apresentada por muitos autores é eliminar variáveis do modelo, através de vários métodos existentes. Dado que o mal condicionamento é causado pela dependência linear aproximada entre algumas das variáveis, se eliminarmos uma delas de cada um dos conjuntos da coluna envolvida estaríamos eliminando o problema de mal condicionamento. Todavia, se delirmos a i-ésima coluna de **X** do modelo estaremos assumindo que desconhecemos o i-ésimo parâmetro de β, β_i seria zero. Se $\beta_i \neq 0$ então o estimador de mínimos quadrados de β fornecido será viciado com o tamanho do vício dependendo do tamanho de β_i. Observamos, aqui, que o estimador de β_j , $j \neq i$, também será viciado, a menos que i-ésima coluna de **X** seja ortogonal as demais colunas. Há técnicas para este procedimento de eliminação de variáveis, entretanto nosso objetivo não é apresentar estas técnicas, pois pressupomos que as variáveis do modelo de regressão disponivéis são altamente importantes, não podendo ser eliminadas. Contudo mostraremos um método de regressão, apropriado a estes casos de multicolinearidade aproximada, onde não é necessario a eliminação das variáveis e nos fornece estimadores mais precisos que os estimadores de mínimos quadrados.

Capítulo 3

Regressão "Ridge"

Notamos que a presença do mal condicionamento de **X** torna grande a variância dos estimadores de mínimos quadrados. O teorema de Gauss Markov garante variância mínima somente dentre os estimadores não viciados mas não garante que esta seja a menor possível, em qualquer situação.

Diante disto, *Arthur Hoerl* (1970), utiliza a regressão "ridge", onde obtém uma variância menor que a dos mínimos quadrados adicionando uma pequena quantidade positiva, ou seja, viciando o estimador da forma[1]:

$$\mathbf{b}(k) = (\mathbf{W}^T\mathbf{W} + k\mathbf{I})^{-1}\mathbf{W}^T\mathbf{y} \text{ , } k \geq 0 \qquad (3.1)$$

A esse tipo de estimador atribuiu-se o nome estimador "ridge", para obte-lo devemos encontrar um valor de k. Mas, qual o valor de k ótimo? Existem várias maneiras de encontrarmos o valor de k. Nosso objetivo neste capítulo é apresentar o método de regressão "ridge" , algumas de suas propriedades e alguns critérios mais utilizados na literatura para obter-se o melhor valor para k.

[1] **W** é a mesma da seção 1.3.

3.1 Estimador "Ridge"

Vamos considerar o modelo de regressão linear múltipla padrão, definido em (1.1). Baseado no mal condicionamento da matriz de variáveis regressoras, um método alternativo de regressão que controla a inflação e a instabilidade geral associado com estimadores de mínimos quadrados é a regressão "ridge", que de uma forma mais geral fornece os seguintes estimadores "ridge":

$$\mathbf{b(K)} = (\mathbf{W^TW+K})^{-1}\mathbf{W^Ty} \tag{3.2}$$

onde, $\mathbf{K}$ é uma matriz diagonal com elementos $(k_1 , \ldots , k_p)$, $k_i \geq 0$ para$\forall i$. São várias as propostas de se obter o estimador "ridge" através de diferentes quantidades positivas adicionadas na diagonal da matriz $\mathbf{W^TW}$, sendo o mais usual esses valores serem todos iguais. Sob esta perspectiva, trabalharemos com a definição (3.1) do estimador "ridge", sem perda de generalidade.

A relação entre estimador "ridge" com o estimador de mínimos quadrados é dada por:

$$\mathbf{b}(k) = (\mathbf{W^TW} + k\mathbf{I})^{-1}\mathbf{W^Ty}$$

denotando $(\mathbf{W^TW} + k\mathbf{I})^{-1} = \mathbf{F}$, teremos

$$\mathbf{b}(k) = \mathbf{FW^Ty} \tag{3.3}$$

e usando (1.2), mas considerando a matriz $\mathbf{W}$ como matriz das variáveis regressoras, teremos

$$\mathbf{b}(k) = \mathbf{FW^TWb} = (\mathbf{W^TW} + k\mathbf{I})^{-1}\mathbf{W^TWb} \tag{3.4}$$

$$= [(\mathbf{W}^T\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{W}) + k(\mathbf{W}^T\mathbf{W})^{-1}]^{-1}\mathbf{b} =$$

$$=[\mathbf{I} + k\,(\mathbf{W}^T\mathbf{W})^{-1}]^{-1}\mathbf{b} \qquad (3.5)$$

denotando $[\mathbf{I} + k\,(\mathbf{W}^T\mathbf{W})^{-1}]^{-1} = \mathbf{Z}$, (3.6)

teremos

$$\mathbf{b}(k) = \mathbf{Z}\mathbf{b}.$$

O valor esperado do estimador é

$$E(\mathbf{b}(k)) = E(\mathbf{Z}\mathbf{b}) = \mathbf{Z}E(\mathbf{b}) = \mathbf{Z}\beta.$$

Desta relação teremos que $\mathbf{b}(k)$ é um estimador viciado se $\mathbf{Z} \neq \mathbf{I}$, sendo que $\mathbf{Z}$ uma matriz que depende de k. Se $\mathbf{Z} = \mathbf{I}$, teremos $k = 0$ o que nos fornece um estimador não viciado, ou seja, o estimador de mínimos quadrados.

3.1.1 Propriedades

Forneceremos a seguir algumas propriedades importantes de $\mathbf{b}(k)$, $\mathbf{F}$ e $\mathbf{Z}$ que serão usadas no decorrer deste trabalho:

P1. Seja $\xi_i(\mathbf{F})$ e $\xi_i(\mathbf{Z})$ os autovalores de $\mathbf{F}$ e $\mathbf{Z}$, respectivamente. Então $\xi_i(\mathbf{F}) = \dfrac{1}{\lambda_i + k}$ e $\xi_i(\mathbf{Z}) = \dfrac{\lambda_i}{\lambda_i + k}$, onde λ_i, i=1,2,...,p, são os autovalores de $\mathbf{W}^T\mathbf{W}$.

Dem.:

Utilizando 1.4 na matriz $\mathbf{W}^T\mathbf{W}$, temos que esta pode ser decomposta em $\mathbf{W}^T\mathbf{W} = \mathbf{V}\Lambda\mathbf{V}^T$ onde $\mathbf{V}$ é matriz de autovetores e $\Lambda = \text{diag}(\lambda_1,\lambda_2,...,\lambda_p)$ são os autovalores

associados aos autovetores, tal que $\lambda_1 > \lambda_2 > \ldots > \lambda_p$. Como os autovalores de $\mathbf{W}^T\mathbf{W}$ são os λ_i's i=1,2,...,p e são decrescentes, o i-ésimo autovalor de $\mathbf{W}^T\mathbf{W}$ é λ_i e o i-ésimo autovetor associado ao autovalor v_i.

Logo, da definição de vetor característica, teremos:

$$\mathbf{W}^T\mathbf{W}\mathbf{v_i} = \lambda_i\mathbf{v_i},\ i = 1,2,\ldots,p \Rightarrow \left| \mathbf{W}^T\mathbf{W} - \lambda\mathbf{I} \right| = 0.$$

Agora, se adicionarmos um a constante à diagonal da matriz $\mathbf{W}^T\mathbf{W}$, isto é, $(\mathbf{W}^T\mathbf{W} + k\mathbf{I})$ e utilizarmos a definição acima, teremos

$$(\mathbf{W}^T\mathbf{W} + k\mathbf{I})\mathbf{v_i} = \mathbf{W}^T\mathbf{W}\,\mathbf{v_i} + k\mathbf{v_i} = \lambda_i\mathbf{v_i} + k\mathbf{v_i} = (\lambda_i + k)\mathbf{v_i} \qquad (3.7)$$

para i = 1,2,...,p, logo

$$\left| \mathbf{W}^T\mathbf{W} - (\lambda+k)\mathbf{I} \right| = 0$$

e, portanto, λ_i+k é autovalor da matriz $(\mathbf{W}^T\mathbf{W} + k\mathbf{I})$.

Invertendo a matriz teremos que o i-ésimo autovalor de $\mathbf{F} = (\mathbf{W}^T\mathbf{W} + k\mathbf{I})^{-1}$ será $(\lambda_i + k)^{-1}$, de fato, se multiplicarmos ambos os membros de (3.7) por $\mathbf{F}$, teremos

$$(\mathbf{W}^T\mathbf{W} + k\mathbf{I})^{-1}(\mathbf{W}^T\mathbf{W} + k\mathbf{I})\mathbf{v_i} = (\mathbf{W}^T\mathbf{W} + k\mathbf{I})^{-1}(\lambda_i + k)\mathbf{v_i}$$

o que implica

$$\mathbf{v_i} = (\lambda_i + k)(\mathbf{W}^T\mathbf{W} + k\mathbf{I})^{-1}\mathbf{v_i};$$

multiplicando a equação por $(\lambda_i + k)^{-1}$,

$$(\lambda_i + k)^{-1}\mathbf{v_i} = (\mathbf{W}^T\mathbf{W} + k\mathbf{I})^{-1}\mathbf{v_i}$$

para i = 1,2,...,p, logo

$$\left|(\mathbf{W}^T\mathbf{W} + k\mathbf{I})^{-1} - (\lambda + k)^{-1}\mathbf{I}\right| = 0$$

e, portanto, $\frac{1}{\lambda_i + k}$ é autovalor da matriz **F** associado ao autovetor $\mathbf{v_i}$, logo $\xi_i(\mathbf{F}) = \frac{1}{\lambda_i + k}$.

Utilizando da igualdade $\mathbf{Z} = \mathbf{F}\mathbf{W}^T\mathbf{W}$ e seguindo o mesmo raciocínio, teremos:

$$\mathbf{Z}\mathbf{v_i} = \mathbf{F}\mathbf{W}^T\mathbf{W}\mathbf{v_i} = \mathbf{F}\,\lambda_i\mathbf{v_i} = \lambda_i\,\mathbf{F}\mathbf{v_i} = \lambda_i(\lambda_i + k)^{-1}\mathbf{v_i}$$

para i = 1,2,...,p, o que implica,

$$\left|\mathbf{Z} - \lambda(\lambda + k)^{-1}\mathbf{I}\right| = 0.$$

Logo, $\xi_i(\mathbf{Z}) = \frac{\lambda_i}{\lambda_i + k}$. ■

P2 **Z** pode ser escrito da forma $\mathbf{Z} = \mathbf{I} - k(\mathbf{W}^T\mathbf{W} + k\mathbf{I})^{-1} = \mathbf{I} - k\mathbf{F}$.

Dem.:

De (3.6) temos que $\mathbf{Z} = [\mathbf{I} + k\,(\mathbf{W}^T\mathbf{W})^{-1}]^{-1}$, logo teremos:

$$\mathbf{Z}\,[\mathbf{I} + k\,(\mathbf{W}^T\mathbf{W})^{-1}] = \mathbf{I},$$

aplicando a propriedade da distributiva teremos, $\mathbf{Z}\mathbf{I} + \mathbf{Z}k\,(\mathbf{W}^T\mathbf{W})^{-1} = \mathbf{I}$, agora

$$\mathbf{Z} = \mathbf{I} - \mathbf{Z}k\,(\mathbf{W}^T\mathbf{W})^{-1} = \mathbf{I} - k\mathbf{Z}(\mathbf{W}^T\mathbf{W})^{-1},$$

sabemos que $\mathbf{F} = \mathbf{Z}(\mathbf{W}^T\mathbf{W})^{-1}$, logo teremos $\mathbf{Z} = \mathbf{I} - k\mathbf{F} = \mathbf{I} - k(\mathbf{W}^T\mathbf{W} + k\mathbf{I})^{-1}$

■

P3 Para $k\neq 0$, $\mathbf{b}(k)$ tem norma menor que $\mathbf{b}$, isto é, $(\mathbf{b}(k))^T(\mathbf{b}(k)) < \mathbf{b}^T\mathbf{b}$.

Dem.:

Por definição $\mathbf{b}(k)=\mathbf{Z}\mathbf{b}$ e usando o fato que $\mathbf{W}^T\mathbf{W}$ e $\mathbf{Z}$ são simétricas positiva definida. De *Riley* (1955), temos que: para uma matriz definida $\mathbf{A}$, $\|\mathbf{A}\| = \lambda_{max}$ e para qualquer matriz $\mathbf{A}$ e vetor $\mathbf{c}$, $\|\mathbf{A}\mathbf{c}\| \leq \|\mathbf{A}\|\,\|\mathbf{c}\|$. Então, a seguinte relação segue:

$$(\mathbf{b}(k))^T(\mathbf{b}(k)) = \|\mathbf{b}(k)\|^2 = \|\mathbf{Z}\mathbf{b}\|^2 \leq \|\mathbf{Z}\|^2\|\mathbf{b}\|^2 = \xi^2_{max}(\mathbf{Z})\mathbf{b}^T\mathbf{b}$$

mas $\xi_{max}(\mathbf{Z}) = \lambda_1/(\lambda_1 + k)$ onde λ_1 é o maior autovalor de $\mathbf{W}^T\mathbf{W}$.

Como $k\neq 0$ temos $\xi_{max}(\mathbf{Z}) < 1$, Então $(\mathbf{b}(k))(\mathbf{b}(k)) < \mathbf{b}^T\mathbf{b}$. ■

3.2 Erro Quadrático Médio Total dos Estimadores "Ridge"

A soma dos erros quadráticos médios de cada estimador a qual denominaremos erro quadrático médio total, é obtida através da distância entre $\mathbf{b}(k)$ e β. Logo, se denotarmos o erro quadrático médio total do estimador por EQMT(k), teremos o EQMT(k) = $E(\mathbf{b}(k) - \beta)^T(\mathbf{b}(k) - \beta)$, deste segue os seguintes resultados:

$$EQMT(k) = E((\mathbf{b}(k) - \beta)^T(\mathbf{b}(k) - \beta)) \qquad (3.7)$$

somando e subtraindo o termo $E(-2\mathbf{b}^T\mathbf{Z}^T\mathbf{Z}\beta + 2\beta^T\mathbf{Z}^T\mathbf{Z}\beta - 2\beta^T\mathbf{Z}^T\beta)$ na expressão (3.7), teremos, facilmente que

$$EQMT(k) = E((\mathbf{b} - \beta)^T\mathbf{Z}^T\mathbf{Z}(\mathbf{b} - \beta)) + (\mathbf{Z}\beta - \beta)^T(\mathbf{Z}\beta - \beta).$$

O segundo termo é a distância ao quadrado de $\mathbf{Z}\beta$ a β. Assim, pode ser considerado como o quadrado do vício. O primeiro termo, veremos mais adiante que é a soma das variâncias (variância total) dos estimadores dos parâmetros. Desenvolvendo cada um dos termos, teremos:

$$EQMT(k)= E[tr(\mathbf{b} - \beta)^T\mathbf{Z}^T\mathbf{Z}(\mathbf{b} - \beta)] + \beta^T(\mathbf{Z} - \mathbf{I})^T(\mathbf{Z} - \mathbf{I})\beta =$$
$$= E(tr\mathbf{Z}^T\mathbf{Z}\,(\mathbf{b} - \beta)(\mathbf{b} - \beta)^T) + \|(\mathbf{Z} - \mathbf{I})\beta\|^2 =$$

usando a propriedade P2 temos que $\mathbf{Z} - \mathbf{I} = -k\mathbf{F}$ onde $\mathbf{F} = (\mathbf{W}^T\mathbf{W} + k\mathbf{I})^{-1}$ logo, teremos a igualdade:

$$EQMT(k)= tr(\mathbf{Z}^T\mathbf{Z}\,E(\mathbf{b} - \beta)(\mathbf{b} - \beta)^T) + \|(-k)\mathbf{F}\beta\|^2 =$$
$$= tr(\mathbf{Z}^T\mathbf{Z}Var(\mathbf{b})) + k^2\|\mathbf{F}\beta\|^2 =$$
$$= \sigma^2\, tr(\mathbf{Z}^T\mathbf{Z}(\mathbf{W}^T\mathbf{W})^{-1}) + k^2\|(\mathbf{W}^T\mathbf{W} + k\mathbf{I})^{-1}\,\beta\|^2 =$$

como $\mathbf{Z} = \mathbf{F}\mathbf{W}^T\mathbf{W}$ implica $\mathbf{Z}(\mathbf{W}^T\mathbf{W})^{-1} = \mathbf{F}$, tem-se,

$$EQMT(k)= \sigma^2\, tr(\mathbf{Z}^T\mathbf{F}) + k^2\, tr[\beta^T(\mathbf{W}^T\mathbf{W} + k\mathbf{I})^{-2}\,\beta] =$$
$$= \sigma^2\, tr(\mathbf{Z}^T\mathbf{F}) + k^2\, tr[\beta^T\beta\,(\mathbf{W}^T\mathbf{W} + k\mathbf{I})^{-2}\,] =$$

fazendo a transformação $\alpha = \mathbf{V}\beta$, onde V é a mesma matriz que na seção 1.4, e tendo que $\beta^T\beta = \alpha^T\mathbf{V}\mathbf{V}^T\alpha = \alpha^T\alpha$, teremos:

$$EQMT(k)= \sigma^2 \, tr(\mathbf{Z}^T\mathbf{F}) + k^2 \, tr[\alpha^T \alpha \, (\mathbf{W}^T\mathbf{W} + k\mathbf{I})^{-2}]$$

utilizando a propriedade P1, temos $\xi_i(\mathbf{F}) = \frac{1}{\lambda_i + k}$ e $\xi_i(\mathbf{Z}) = \frac{\lambda_i}{\lambda_i + k}$, como **Z** e **F** são matrizes diagonais teremos $\xi_i(\mathbf{ZF}) = \frac{\lambda_i}{(\lambda_i + k)^2}$, logo segue que:

$$EQMT(k)= \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i + k)^2} =$$

$$= \gamma_1(k) + \gamma_2(k) \qquad (3.8)$$

onde $\gamma_1(k)$ e $\gamma_2(k)$ são, respectivamente, a variância total e vício-quadrado do estimador "ridge". Outra maneira para verificar que $\gamma_1(k)$ corresponde a variância total é dada em termos do estimador de mínimos quadrados

$$\mathbf{b}(k) = \mathbf{Zb},$$

então,

$$Var\,(\mathbf{b}(k)) = Var\,(\mathbf{Zb}) =$$

$$= \mathbf{Z}\, Var(\mathbf{b})\mathbf{Z}^T$$

e, utilizando o resultado da variância do estimador de mínimos quadrados, teremos

$$Var\,(\mathbf{b}(k)) = \sigma^2\, \mathbf{Z}\, (\mathbf{W}^T\mathbf{W})^{-1}\, \mathbf{Z}^T. \qquad (3.9)$$

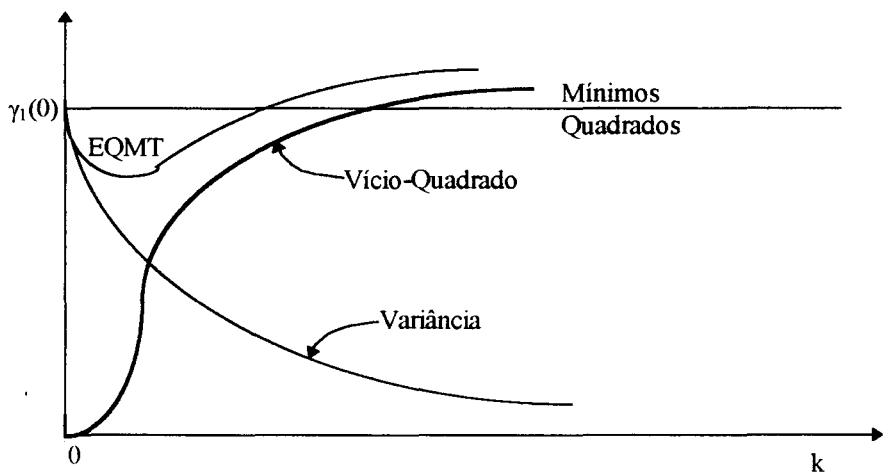
A soma da variância de todos os $b_i(k)$'s é a soma dos elementos da diagonal de (3.9), ou seja, é a soma dos autovalores da matriz $\mathbf{Z}\, (\mathbf{W}^T\mathbf{W})^{-1}\, \mathbf{Z}^T$, já vimos que $\xi_i(\mathbf{Z})$

$= \frac{\lambda_i}{\lambda_i + k}$ e que os autovalores de $(\mathbf{W}^T\mathbf{W})^{-1}$ são λ_i^{-1} i=1,...,p, logo os autovalores da matriz $\mathbf{Z}(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{Z}^T$ são $(\frac{\lambda_i}{\lambda_i + k})^2\lambda_i^{-1}$ i=1,...,p, portanto a soma da variância total será dada por:

$$VART(\mathbf{b}(k)) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2}$$

A figura abaixo mostra o esboço do comportamento das funções $\gamma_1(k)$, $\gamma_2(k)$ e a soma de ambas.

Figura 1.1: Variância, o vício e a soma de ambos, EQMT, como função de k



Verificamos que quando k = 0 o estimador "ridge", dado por (3.1), é igual ao estimador de mínimos quadrados, neste caso o vício-quadrado e a variância de **b**(0) são, respectivamente, iguais a $\gamma_2(0) = 0$ e $\gamma_1(0) = \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}$. Como a matriz de regressão é mal condicionada, temos $\lambda_i \rightarrow 0$, pois a variância do estimador de mínimos quadrados é muito grande. A medida que cresce o valor de k aumenta o valor do vício e diminui a variância, isto nos leva a estimadores viciados, no entanto, com variância menor que a dos mínimos quadrados, de modo que o vício tenda a $\beta^T\beta$ e a variância a zero. Como indicado pelo gráfico, a soma de $\gamma_1(k)$ e $\gamma_2(k)$ resulta na soma dos erros quadráticos médio, observamos

que quando k = 0 o erro quadrático médio é igual a variância do estimador de mínimos quadrados, a medida que k cresce o EQMT diminui para um ponto mínimo global, voltando, novamente, a aumentar quando k é muito grande e chegando a ser maior que a variância do estimador de mínimos quadrados. Podemos dizer que, graficamente, há valores de k nos quais a soma dos erros quadráticos médio de $\mathbf{b}(k)$ é menor do que $\mathbf{b}$. Estas afirmações nos conduz à conclusão que é possível encontrar $k > 0$, com vício pequeno, que reduz, substancialmente, a variância e melhorando desta maneira, o erro quadrático médio do estimador e do predito. Isto é comprovado na seção 3.4 onde provamos, algebricamente, que sempre existe um valor de k não negativo tal que o erro quadrático médio do estimador "ridge" é menor que o erro quadrático médio do estimador de mínimos quadrados.

3.3 Erro quadrático Médio Total do Predito

A soma dos erros quadráticos médio do predito, como em 3.2, é obtido da distância entre $\hat{\mathbf{y}}^*$ e $\mathbf{y}$, onde $\hat{\mathbf{y}}^*$ é a estimativa da variável resposta do modelo "ridge". Aqui, denotaremos o erro quadrático médio total do predito por EQMTP e teremos

$$EQMTP(k) = E(\hat{\mathbf{y}}^* - E(\hat{\mathbf{y}}))^T(\hat{\mathbf{y}}^* - E(\hat{\mathbf{y}})).$$

Desenvolvendo esta expressão,

$$EQMTP(k) = E(\mathbf{W}\mathbf{b}(k) - \mathbf{W}\beta)^T(\mathbf{W}\mathbf{b}(k) - \mathbf{W}\beta) =$$

$$= E(\mathbf{b}(k) - \beta)^T\mathbf{W}^T\mathbf{W}(\mathbf{b}(k) - \beta) =$$

$$= E[tr(\mathbf{b}(k) - \beta)^T\mathbf{W}^T\mathbf{W}(\mathbf{b}(k) - \beta)] =$$

$$= E[tr\mathbf{W}^T\mathbf{W}(\mathbf{b}(k) - \beta)(\mathbf{b}(k) - \beta)^T] =$$

$$= tr[(\mathbf{W}^T\mathbf{W})E(\mathbf{b}(k) - \beta)(\mathbf{b}(k) - \beta)^T] =$$

$$= tr[(\mathbf{W}^T\mathbf{W})EQMT]$$

já vimos que,

$EQMT(k) = E((\mathbf{b}(k) - \beta)^T(\mathbf{b}(k) - \beta)) = \sigma^2 \, tr(Z^TF) + k^2 \, tr[\alpha^T \alpha \, (\mathbf{W}^T\mathbf{W} + k\mathbf{I})^{-2}] =$

$= \sum_{i=1}^{p} \frac{\sigma^2\lambda_i + k^2\alpha_i^2}{(\lambda_i + k)^2}$, logo a forma geral dos elementos da diagonal do EQMT é dado por $\frac{\sigma^2\lambda_i + k^2\alpha_i^2}{(\lambda_i + k)^2}$; multiplicando-o pelo elemento da diagonal de $\mathbf{W}^T\mathbf{W}$, teremos o resultado:

$$EQMTP(k) = \sum_{i=1}^{p} \frac{\sigma^2\lambda_i^2 + k^2\lambda_i\alpha_i^2}{(\lambda_i + k)^2}$$

Observe que o erro quadrático médio total do estimador difere do erro quadrático médio total do predito pela multiplicação do autovalor da matriz $\mathbf{W}^T\mathbf{W}$ na diagonal principal do EQMT, não alterando, substancialmente, suas propriedades.

3.4 Teoremas sobre a função Erro Quadrático Médio Total

Teorema 1: A variância total $\gamma_1(k)$ é uma função contínua e monótona decrescente.

Dem.:

Sabemos que $\gamma_1(k) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2}$.

Para $k > 0$, temos $k < k + \delta$, $\forall \, \delta > 0$

$$\gamma_1(k) - \gamma_1(k + \delta) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} - \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k + \delta)^2} =$$

$$= \sigma^2 [\sum_{i=1}^{p} (\frac{\lambda_i}{(\lambda_i + k)^2} - \frac{\lambda_i}{(\lambda_i + k + \delta)^2})] =$$

dado que $\lambda_i > 0$ teremos que $\frac{\lambda_i}{(\lambda_i + k)^2} > \frac{\lambda_i}{(\lambda_i + k + \delta)^2}$,

portanto, $\frac{\lambda_i}{(\lambda_i + k)^2} - \frac{\lambda_i}{(\lambda_i + k + \delta)^2} > 0$

logo, teremos: $\gamma_1(k) > \gamma_1(k + \delta)$

Portanto, $\gamma_1(k)$ é função monótona decrescente de k.

A função $\gamma_1(k)$ é contínua, pois trata-se da soma de racionais nas quais o denominador nunca se anula, logo $\forall\ k_0 > 0$,$\exists$ limite $\gamma_1(k)$ e $\lim_{k \to k_0} \gamma_1(k) = \gamma_1(k_0)$ ■

Corolário 1.1 A primeira derivada com respeito a k da variância total $\gamma_1'(k)$, tende a $-\infty$ quando $k \to 0^+$ e $\lambda_p \to 0$.

Dem.:

Derivando $\gamma_1(k)$, teremos $\gamma_1'(k) = -2\sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^3}$ que é uma função contínua, cujo limite quando $k \to 0^+$ é igual a $\gamma_1'(0) = -2\sigma^2 \sum_{i=1}^{p} \frac{1}{(\lambda_i)^2}$ e considerando $\lambda_p \to 0$ teremos que $\gamma_1'(0) \to -\infty$ ■

Teorema 2: O vício quadrado $\gamma_2(k)$ é uma função contínua e monótona crescente de k.

Dem.:

De (3.8) temos

$$\gamma_2(k) = k^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i + k)^2}, \text{ onde } \alpha = \mathbf{V}\beta$$

Dado $\lambda_i > 0$ $\forall i=1,2,...,p$ e $k \geq 0$, os elementos $(\lambda_i + k)^{-2}$ para $i = 1,2,...,p$, nunca se anulam, claramente são funções continuas, por outro lado k também é continuo. O produto de funções continuas são continuas, logo $\gamma_2(k)$ é contínua para $k \geq 0$, isto é, $\forall\ k_0 > 0$ $\exists$ limite $\gamma_2(k)$ e $\lim_{k \to k_0} \gamma_2(k) = \gamma_2(k_0)$

Para $k>0$, podemos rescrever $\gamma_2(k)$ como:

$$\gamma_2(k) = \sum_{i=1}^{p} \frac{\alpha_i^2}{(1 + \lambda_i / k)^2},$$

Como $\lambda_i > 0$ $\forall i$, a função λ_i / k é monótona decrescente quando k cresce, assim o termo $\dfrac{\alpha_i^2}{(1 + \lambda_i / k)^2}$ é monótona crescente. Tendo que $\gamma_2(k)$ é uma função monótona crescente. ■

Corolário 2.1: O vício quadrado $\gamma_2(k)$ aproxima-se de $\beta^T\beta$ como um limite superior.

Dem.:

$$\lim_{k \to \infty} \gamma_2(k) = \sum_i \alpha_i^2 = \alpha^T\alpha = \beta^T\mathbf{V}^T\mathbf{V}\beta = \beta^T\beta.$$ ■

Corolário 2.2: A derivada $\gamma_2'(k)$ tende a zero quando $k \to 0^+$.

Dem.:

Do teorema 2, temos:

$$\gamma_2(k) = k^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i + k)^2} \Rightarrow \gamma_2'(k) = 2k \sum_{i=1}^{p} \frac{\lambda_i \alpha_i^2}{(\lambda_i + k)^3}.$$

Cada termo $\dfrac{2k\lambda_i\alpha_i^2}{(\lambda_i + k)^3}$ é uma função contínua.

Logo, $\lim_{k \to 0} \gamma_2'(k) = \gamma_2'(0) = 0$. ■

Notamos ainda, que os valores das derivadas destas funções γ_1 e γ_2 no limite da origem tem grande significado.

$$\lim_{k\to 0^+}\left(\frac{d\gamma_1}{dk}\right) = -2\sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_i^2} \tag{3.10}$$

$$\lim_{k\to 0^+}\left(\frac{d\gamma_2}{dk}\right) = 0 \tag{3.11}$$

Vemos que γ_1 (k) tem uma derivada negativa quando $k \to 0^+, \frac{\partial\gamma_1}{\partial k} \to -2p\sigma^2$ quando a uma matriz $\mathbf{W}^T\mathbf{W}$ é ortogonal e aproximando-se de $-\infty$ quando $\mathbf{W}^T\mathbf{W}$ passa a ser mal condicionada e $\lambda_p \to 0$.

Por outro lado, quando $k \to 0^+$ (3.11) mostra que γ_2 (k) é zero.

Teorema 3: (Teorema da Existência). Existe um $k > 0$ tal que $EQMT(k) < EQMT(0) = \sigma^2 \sum_i \frac{1}{\lambda_i}$

Dem.:

$$\text{De (3.8) temos } EQMT(k) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i + k)^2}$$

derivando a função EQMT(k), teremos

$$\frac{dEQMT(k)}{dk} = \frac{d\gamma_1(k)}{dk} + \frac{d\gamma_2(k)}{dk} = -2\sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^3} + 2k \sum_{i=1}^{p} \frac{\lambda_i \alpha_i^2}{(\lambda_i + k)^3}$$

Primeiro observamos que γ_1 (0) = $\Sigma 1/\lambda_i$ e γ_2 (0) = 0 $\Rightarrow$ E(EQMT(0)) = $\sigma^2 \Sigma 1/\lambda_i$.

Nos teoremas 1 e 2 provou-se que $\gamma_1(k)$ e $\gamma_2(k)$ são funções monótonas decrescente e crescente, respectivamente. Suas primeiras derivadas são sempre não negativa e não positivas, respectivamente. Desta forma, para provar o teorema, é somente necessário provar que existe k>0 tal que $\frac{dEQMT(k)}{dk} < 0$.

$$\text{Logo, } \frac{dEQMT(k)}{dk} = -2\sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^3} + 2k \sum_{i=1}^{p} \frac{\lambda_i \alpha_i^2}{(\lambda_i + k)^3} =$$

$$= \sum_i \frac{-2\sigma^2 \lambda_i + 2k\lambda_i \alpha_i^2}{(\lambda_i + k)^3} = \sum_i \frac{2\lambda_i(-\sigma^2 + k\alpha_i^2)}{(\lambda_i + k)^3} < 0$$

uma condição para esta expressão ser negativa é que:

$$-\sigma^2 + k\alpha_i^2 < 0, \text{ logo } k \leq \frac{\sigma^2}{\alpha_i^2} \quad \forall\, i = 1,2, \ldots, p$$

Então temos que $k \leq \frac{\sigma^2}{\alpha_{max}^2}$, portanto, existe um k tal que EQMT(k) < EQMT(0)

■

Observação: As Propriedades de EQMT = γ_1 (k) + γ_2 (k) mostram que há um valor de k>0 tal que a função erro quadrático médio total tende ao mínimo.

3.5 Métodos de Escolha do k "Ótimo".

Existem na literatura várias propostas de escolha para k, neste trabalho nos deteremos em algumas delas.

Método I: Introduzido por *Hoerl e Kennard* (1970), que sugere um estimador "ridge" de modo que o EQMT dos estimadores seja mínimo. Considerando a função obtida na seção 3.2, *Hoerl e Kennard* mostra que o EQMT é mínimo quando $k_i=\frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}$.

De fato:

$$EQMT(k_i)=\sum_{i=1}^{p}\frac{\lambda_i\sigma^2+k_i^2\alpha_i^2}{(\lambda_i+k_i)^2}$$

onde, o EQMT é dado utilizando a forma geral do "ridge" e considerando cada termo da soma, teremos:

$$f(k_i)=\frac{\lambda_i\sigma^2+k_i^2\alpha_i^2}{(\lambda_i+k_i)^2} \tag{3.12}$$

derivando f em relação a k_i, obtém-se:

$$f'(k_i)=\frac{2k\alpha_i^2(\lambda_i+k_i)^2-2(\lambda_i+k_i)(\lambda_i\sigma^2+k_i^2\alpha_i^2)}{(\lambda_i+k_i)^4}$$

$$f'(k_i)=0 \Leftrightarrow [2k\alpha_i^2(\lambda_i+k_i)-2(\lambda_i\sigma^2+k_i^2\alpha_i^2)](\lambda_i+k_i)=0$$

$$\Leftrightarrow 2k_i\alpha_i^2(\lambda_i+k_i)-2(\lambda_i\sigma^2+k^2\alpha_i^2)=0, \text{ pois } (\lambda_i+k_i)>0$$

então, a expressão acima será nula, se $k_i=\frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}$.

Logo, $k_i=\frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}$, i=1,...,p são pontos que minimizam a função EQMT. ■

Método II: Este método é introduzido por *Hoerl, Kennard e Baldwin* (1975). Eles consideram uma combinação dos k_i's , do método I, em um único valor de k, onde este é obtido da média harmônica dos k_i's. Seja k_h a média harmônica, então teremos:

$$\frac{1}{k} = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{k_i} = \frac{1}{p}\sum_{i=1}^{p}\frac{\alpha_i^2}{\sigma^2} = \frac{1}{p\sigma^2}\sum_{i=1}^{p}\alpha_i^2$$

$$k_h = \frac{p\hat{\sigma}^2}{\sum_{i=1}^{p}\hat{\alpha}_i^2}$$

Constatamos que a média aritmética não é uma boa escolha, pois pequenos valores de α_i produzem valores muito grandes para k resultando em um vício maior ainda.

Os dois métodos apresentados dependem dos parâmetros de σ^2 e α_i , i=1,2, ..., p; na prática substituímos os valores desses parâmetros por suas estimativas, ou seja, QME para σ^2 o estimador de mínimos quadrados para α_i, i=1,2,...,p.

Há autores[2] que sugerem um processo iterativo dos métodos I e II. Nestes processos o valor inicial de k, digamos k_j, é obtido utilizando o estimador de mínimos quadrados para α_j, obtendo-se o estimador "ridge", $\alpha_j(k_j)$. O segundo valor de k, k_{j+1}, é calculado utilizando $\alpha_j(k_j)$ como estimativa de α_j. Este processo é repetido para j = 1,2,3 ...., até que a diferença entre k_{j+1} e k_j seja menor que um valor δ, digamos $\delta = 10^{-4}$. Alguns estudos comprovam que este processo melhora EQMT, isto é, nos fornece um valor menor para este quando comparado aos métodos I e II.

Entretanto, nem sempre conseguimos a convergência. *Gibbons* (1981) afirma que esta convergência deve ser obtida até 30 interações. No desenvolvimento deste trabalho, em alguns dos nossos estudos esta convergência não foi obtida, por isso optamos em não incluí-la na simulação.

---

[2] *Hoerl e Kennard* (1976) sugere processo iterativo do método II.

Método III: Este método foi fornecido por *Hemmerle* (1975). Baseado no processo iterativo do método I, propõe um processo não iterativo, entretanto, que nos dá uma solução aproximada da obtida por Hoerl. Esta solução dependerá somente de uma condição de convergência/divergência.

Em particular, seja:

$$e_i = \frac{\hat{\sigma}^2}{\lambda_i \hat{\alpha}_i^2}$$

e

$$e_i^* = \frac{1 - e_i - \sqrt{1 - 4e_i}}{2e_i}, \text{ para } e_i \leq \frac{1}{4}$$

então,

$$\hat{\alpha}_i^* = \begin{cases} 0 \ , \text{ se } e_i > \dfrac{1}{4} \\ \dfrac{\hat{\alpha}_i}{1 + e_i^*} \ , \text{ se } e_i \leq \dfrac{1}{4} \end{cases}$$

assim,

$$\mathbf{b}^*(k) = \mathbf{V}^T \hat{\alpha}^*$$

Método IV: Este método é utilizado em *Lee e Campbell* (1985), conhecido por método de *Newton Raphson*, minimiza com respeito a k, a função EQMT dada na seção 3.2. Em virtude disso, um algoritmo iterativo para obter o parâmetro "ridge" é fornecido como segue:

Passo1: $k^{(0)} = 0$ e $i = 0$

Passo2: Calcule $k^{(i+1)}$ de

$$k^{(i+1)} = k^{(i)} + \frac{f'(k^{(i)})}{f''(k^{(i)})} \qquad (3.13)$$

onde, f' e f'' são, respectivamente, a primeira e segunda derivada de f, definida em (3.12).

Passo3: Se $| k^{(i+1)} - k^{(i)} | < \delta$ para algum dado $\delta>0$, pare. Caso contrário, considere i:= i + 1 e vá para o passo 2.

A equação (3.13) convergirá para o primeiro mínimo local de f cujo o valor de k está próximo da origem fornecendo, assim, o vício muito pequeno. Observe que tanto este método como método I minimizam a função EQMT, mas, no primeiro caso, obtemos o mínimo global, enquanto que neste obtemos o mínimo local. Tendo como preocupação fornecer um estimador com menor vício possível obtendo ainda uma variância, apesar de grande, menor que o dos mínimos quadrados.

Método V: *Lawless e Wang* (1976), basearam-se no artigo de Efron e Morris com o objetivo de atingir melhores resultados que os obtidos pelo método II , sugeriu o seguinte valor para k:

$$k=\frac{p\hat{\sigma}^2}{\sum_{i=1}^{p}\lambda_i\hat{\alpha}_i^2}$$

Método VI: *Mallows* (1973), generaliza sua estatística C_p para o caso da regressão "ridge". Utilizando da estimativa do EQMTP dado por $\frac{1}{\sigma^2}E\{\|\mathbf{y}^*-E(\mathbf{y})\|^2\}$, Mallows obtém a estatística dada por:

$$C(k_i) = 1 + \sum_{i=1}^{p}\frac{\lambda_i^2 - k_i^2 + \lambda_i k_i^2\hat{\alpha}_i^2 / \hat{\sigma}^2}{(\lambda_i + k_i)^2} \qquad (3.14)$$

Minimizando esta estatística obtemos o valor de k. Além disto, derivando a função acima teremos que k:

$$k_i = \frac{\lambda_i}{\frac{\lambda_i\hat{\alpha}_i^2}{\hat{\sigma}^2} - 1}.$$

O estimador ajustado será dado por:

$$b_i^* = \frac{\lambda_i b_i}{\lambda_i + \frac{\lambda_i}{T_i}}, \text{ onde } T_i = \frac{\lambda_i b_i}{\hat{\sigma}^2} - 1$$

ou

$$b_i^* = (1 - \frac{\hat{\sigma}^2}{\lambda_i b_i^2}) b_i.$$

Capítulo 4

Simulação

Neste capítulo faremos o estudo comparativo de alguns dos métodos, de obtenção de k, propostos no capítulo anterior. Este estudo será feito através de simulações, onde desta tiraremos conclusões sobre o desempenho de cada método sobre uma classe de todos problemas de multicolinearidade aproximada na regressão.

Para isso geramos um conjunto de dados com correlações (c^2) pré-determinadas, denominar-la-emos correlação teórica. Para cada uma das correlações geradas, determinaremos diferentes desvios padrões e para cada um destes pares geraremos 1000 modelos de regressão com erros normalmente distribuídos com média zero e o desvio padrão determinado anteriormente.

Desse conjunto de procedimentos tiraremos conclusões sobre, em qual das situações cada método é melhor indicado, levando em consideração a correlação, o desvio padrão, a variância, o vício, o EQMT e o EQMTP.

Os procedimentos das simulações foram baseados, entre outros, em *Lawless e Wang* (1976). Poderemos vê-los com mais detalhes nas próximas seções, os resultados se encontram em 4.2. Em 4.3 mostraremos um exemplo, onde o conjunto de dados são os índices mensais das bolsas de São Paulo e Rio de Janeiro, neste poderemos ver o comportamento de alguns métodos comparado com o método de mínimos quadrados. Para finalizar, teremos a conclusão do trabalho.

4.1 Geração dos dados

Vamos considerar um modelo de regressão linear múltipla com três variáveis regressoras e um total de quinze observações. Essas variáveis regressoras terão coeficientes de correlação iguais a c^2. Para tanto, serão geradas da seguinte forma:

$$x_{ij} = (1 - c^2)^{1/2} z_{ij} + cz_{i4}\ ,\ i = 1,2, \ldots , 15;\ j = 1,2,3.$$

onde $\mathbf{z}_{i1}$, $\mathbf{z}_{i2}$, $\mathbf{z}_{i3}$, $\mathbf{z}_{i4}$ são números pseudo-aleatórios independentes com distribuição normal padrão e c^2 é o coeficiente de correlação pré determinado, consideraremos quatro diferentes conjuntos de valores para c^2 .80, .90, .95, .99.

Se considerarmos duas variáveis regressoras x_{ij} e x_{ik} teremos:

$$\text{corr}(x_{ij},x_{ik}) = \frac{E(x_{ij}x_{ik}) - E(x_{ij})E(x_{ik})}{\sigma(x_{ij})\sigma(x_{ik})}$$

$$\sigma(x_{ij}) = \sqrt{\text{Var}(x_{ij})} = \sqrt{\text{Var}[(1 - c^2)^{\frac{1}{2}} z_{ij} + c\mathbf{z}_{i4}]} =$$

$$= \sqrt{(1 - c^2)\text{Var}z_{ij} + c^2\text{Var}\mathbf{z}_{i4}} = \sqrt{(1 - c^2) + c^2} = 1$$

$$\sigma(x_{ij}) = \sigma(x_{ik}) = 1$$

$$E(x_{ij}) = E(x_{ik}) = 0.$$

Então, $\text{corr}(x_{ij},x_{ik}) = E(x_{ij}x_{ik})$ e

$$E(x_{ij}x_{ik}) = E\{[(1 - c^2)^{1/2}z_{ij} + cz_{i4}][\ (1 - c^2)^{1/2}z_{ik} + cz_{i4}]\} =$$

$$= E[(1 - c^2)z_{ij}z_{ik} + (1 - c^2)^{1/2}cz_{ij}z_{i4} + (1 - c^2)^{1/2}cz_{ik}\,z_{i4} + c^2 z_{i4}\,z_{i4}]=$$

$$= (1 - c^2)E(z_{ij}z_{ik}) + (1 - c^2)^{1/2}cE(z_{ij}z_{i4}) + (1 - c^2)^{1/2}cE(z_{ik}\,z_{i4}) + c^2\,E(z_{i4}\,z_{i4}) =$$

como z_{ij} e z_{ik} são independentes $j \neq k$, então $E(z_{ij}z_{ik}) = E(z_{ij})E(z_{ik}) = 0$

$$= c^2\,E(z_{i4}\,z_{i4}) = c^2E(z_{i4}^2) = c^2Var(z_{i4}).$$

Agora, como z_{i4} tem distribuição normal padrão, teremos

$$corr(x_j,x_k) = E(x_jx_k) = c^2. \qquad \blacksquare$$

Desta forma a matriz de correlação de **X** terá estrutura de correlação intraclasses[1], isto é:

$$Corr(\mathbf{X}) = \begin{pmatrix} 1 & c^2 & c^2 \\ c^2 & 1 & c^2 \\ c^2 & c^2 & 1 \end{pmatrix}.$$

Construída a matriz X de regressoras, antes de gerarmos os modelos de regressão padronizar-la-emos, centrando e escalonando como definido no capítulo I, passando a chamá-la de W, então o modelo de regressão $y = W\beta + \varepsilon$ será transformado para uma forma ortogonal $y = Z\alpha + \varepsilon$ pela transformação $Z = XV^T$ e $\alpha = V\beta$, onde $X^TX = V^T\Lambda V$, $\Lambda = diag(\lambda_1,\lambda_2,...,\lambda_p)$ são os autovalores de X^TX e V matriz de autovetores; computacionalmente faremos isto utilizando a decomposição de valores singulares (DVS).

Com isso, construiremos o modelo transformado através dos procedimentos seguintes.

---

[1] Mais detalhes desta estrutura pode ser vista em *McDonald* (1979).

4.1.1 Vetor de Coeficiente das Variáveis Regressoras

O coeficiente das regressoras é obtido escolhendo um valor r^2 no intervalo (4,4900), este valor será o comprimento, ao quadrado, do vetor de coeficientes da regressão. Em alguns trabalhos este número é considerado simplesmente um. Dado este valor de r^2, p números aleatórios são escolhidos da distribuição uniforme em $(-1,1)$, a qual chamaremos u. Então calculamos $r_u^2 = \sum_{i=1}^{p} u_i^2$. O coeficiente de regressão é então da forma $\alpha_i = \frac{r}{r_u} u_i$, tal que:

$$\| \alpha \|^2 = \sum_i \alpha_i^2 = \sum_i (\frac{r}{r_u})^2 u_i^2 = r^2$$

4.1.2 Erro

O erro será gerado através da distribuição normal, com média zero e desvios padrões iguais a: 0.55, 0.7, 0.89, 1.0, 3.0 e 5.0. Os desvios padrões foram escolhidos de forma aleatória e os erros são gerados através de procedimento do SAS-IML.

4.1.3 Estimação

Construído o modelo poderemos estimar seus parâmetros, mas, antes disto, calculamos o índice de condição e o fator de inflação da variância, para analisarmos o grau de multicolinearidade. Feito isto, calculamos: o estimador de mínimos quadrados, $\hat{\alpha}$, a variância estimada, $\hat{\sigma}^2$, e o erro quadrático médio total dos estimadores, EQMT(0).

Em seguida, encontraremos os valores de k dos métodos fornecidos na seção (3.6). Para cada valor de k encontraremos os respectivos estimadores "ridge" e calcularemos a variância, o vício, o erro quadrático médio de cada estimador "ridge" e do predito.

4.1.4 Replicação

Para cada tripla, coeficiente de correlação, desvio padrão e r^2, 1000 amostras de tamanho 15 são geradas. As 1000 diferentes amostras são obtidas gerando diferentes erros. Para cada um desses m modelos fazemos as respectivas estimações e, desta maneira, poderemos medir o fator de inflação da variância e o número de condição e comparar os diferentes estimadores "ridge" através da análise do EQMT, do EQMTP, da variância e do vício.

4.2 Resultados

Os resultados da simulação são mostrados nos quadros e gráficos ao longo desta seção. Primeiramente, mostraremos a relação de cada correlação teórica, considerando as correlações amostrais, com número de condição e o VIF.

Quadro 4.1 Diagnóstico de multicolinearidade considerando a correlação teórica e a respectiva correlação amostral.

c^2 **teórico**	$\hat{c}_{12}^2$	$\hat{c}_{13}^2$	$\hat{c}_{23}^2$	**número de condição** η	**VIF**
0.8	0.69	0.56	0.55	7.2023349	2.11086 2.08247 1.57034
0.9	0.89	0.88	0.87	25.503808	6.11800 5.90983 5.56580
0.95	0.94	0.96	0.95	84.010961	16.2517 11.1043 18.1110
0.99	0.99	0.99	0.99	5919.397	145.75 1116.71 915.55

Observando o quadro, vemos que as correlações amostrais estão bem próximas das correlações teóricas, sendo exatamente igual no caso de $c^2 = 0.99$. A medida que aumentamos a correlação entre as variáveis regressoras aumentam-se os valores do número de condição e do VIF.

Pela teoria apresentada na seção 2.4.2, o problema de mal condicionamento é identificado quando o número de condição é maior que 100. Em vista disto, o mal condicionamento ocorreu quando o coeficiente de correlação teórica é igual a 0.99. Já no caso do VIF pela seção 2.4.1 teríamos evidências de mal condicionamento quando o coeficiente de correlação assume os valores 0.95 e 0.99.

Não nos preocupamos com a diferença dos resultados obtidos pelos diagnósticos do número de condição e do VIF, pois esses pontos de corte ainda são fatores em discussão no estudo de multicolinearidade. No entanto, consideraremos a presença de mal condicionamento da matriz, quando assumirmos o coeficiente de correlação entre as variáveis regressoras, a partir de 0.95.

As comparações do estimador de mínimos quadrados com os estimadores "ridge", podem ser vistas no quadro 4.2. Neste quadro, temos a freqüência que EQMT dos estimadores "ridge" será maior que do estimador de mínimos quadrados. Vemos na primeira linha do quadro 4.2 os símbolos que identificam: os coeficientes de correlação, os desvios padrões e cada método de obtenção de k, em seguidas seus respectivos valores .

Quadro 4.2 Número de vezes que EQMT(0) é menor que EQMT(k).

c^2	σ	I	II	III	IV	V	VI
0.8	**0.55**	0	0	0	0	25	1
	0.7	0	0	0	0	336	0
	0.89	0	0	0	0	1000	0
	1.0	0	0	0	0	0	0
	3.0	0	0	0	0	585	0
	5.0	0	0	0	0	259	0
0.9	**0.55**	0	0	0	0	88	0
	0.7	0	0	0	0	103	0
	0.89	0	0	0	0	433	0
	1.0	0	0	0	0	0	0
	3.0	0	0	0	0	136	0
	5.0	0	0	0	0	0	0
	0.55	0	0	0	0	169	0
	0.7	0	0	0	0	304	0

0.95	**0.89**	0	0	0	0	0	0
	1.0	0	0	2	0	385	0
	3.0	0	1	0	0	99	0
	5.0	0	2	0	0	225	0
0.99	**0.55**	0	4	0	0	50	0
	0.7	0	4	0	4	57	0
	0.89	0	28	0	2	195	0
	1.0	0	0	0	1	76	0
	3.0	0	3	0	4	68	0
	5.0	0	2	0	0	75	0

No quadro 4.2 os métodos I, II, III, IV e VI têm, quase sempre, 100% dos casos EQMT(k) < EQMT(0), logo o valor zero está indicando que em todos os casos dos coeficientes de correlação e desvio padrão os métodos mencionados não obtiveram seus EQMT(k) maior que EQMT(0). Já no método V o número de ocorrências de EQMT(0) < EQMT(k) é grande, porém diminui com o aumento da correlação; podendo ver que quando $c^2 = 0.8$ encontramos uma frequência de 336, 585 e até 1000. Quando $c^2 = 0.99$ esta frequência é sempre menor que 200, tendo em média uma frequência de 87, o equivalente a 8.7% de casos com EQMT(0) < EQMT(k). Assim, a presença do mal condicionamento faz do método V de estimação "ridge", em média, ter EQMT(0) maior que dos estimadores "ridge".

Importante salientar que os métodos I, II, III, IV e VI com diferentes desvios padrões não afetam a performance dos estimadores "ridge". Mesmo com c^2 baixo, ou seja, com um mal condicionamento da matriz X não muito acentuado, os estimadores "ridge" se comportam de forma melhor que os ordinários de mínimos quadrados, em termos de seus EQMT.

Além das análises comparando todos os métodos "ridge" com os mínimos quadrados também faremos comparaçõesentre os métodos. Assim, no decorrer deste trabalho, veremos quadros como 4.3, onde consideraremos o caso de correlação teórica 0.8 com todos desvios padrões. Nele as linhas correspondem as frequências em que o EQMT(k) de um determinado método é menor que dos outros, considerando $c^2 = 0.8$ e σ especificado. As colunas correspondem às frequências que o EQMT(k) de um determinado método é maior que dos outros. O total das linhas e colunas são as somas das frequências e por estas poderemos analisar o quanto cada método é melhor ou pior que os outros. Por exemplo, na primeira linha o método I tem um total de frequência igual a 5000, isto é, EQMT(k) é sempre menor que os método II, III, IV, V e VI quando $c^2 = 0.8$ e $\sigma = 0.55$ e a coluna de I

tem total zero que corresponde dizer que este método nunca tem EQMT(k) maior que qualquer outro método quando $c^2 = 0.8$.

Os quadros 4.4, 4.5 e 4.6 levam em conta as correlações 0.9, 0.95 e .099, respectivamente.

Quadro 4.3 Número de vezes em que EQMT(k) de cada um dos métodos da linha é menor que os da coluna, com coeficiente correlação teórico 0.8.

σ		I	II	III	IV	V	VI	Total
0.55	I		1000	1000	1000	1000	1000	5000
	II	0		1000	0	907	972	2879
	III	0	0		0	25	1	26
	IV	0	1000	1000		1000	1000	4000
	V	0	93	975	0		769	1837
	VI	0	28	999	0	231		1258
0.7	I		1000	1000	1000	1000	1000	5000
	II	0		972	0	856	586	2414
	III	0	28		2	366	6	402
	IV	0	1000	998		998	1000	3996
	V	0	144	634	2		422	1202
	VI	0	414	994	0	578		1986
0.89	I		1000	1000	1000	1000	1000	5000
	II	0		981	0	1000	8	1989
	III	0	19		5	1000	10	1034
	IV	0	1000	995		1000	1000	3995
	V	0	0	0	0		0	0
	VI	0	992	990	0	1000		2982
1.0	I		1000	1000	1000	1000	1000	5000
	II	0		998	0	683	1000	2681
	III	0	2		2	2	8	14
	IV	0	1000	998		985	1000	3983
	V	0	317	998	15		991	2321
	VI	0	0	992	0	9		1001
3.0	I		1000	1000	1000	1000	1000	5000
	II	0		993	0	962	497	2452
	III	0	7		0	595	0	602
	IV	0	1000	1000		1000	1000	4000
	V	0	38	405	0		199	642
	VI	0	503	1000	0	801		2304
5.0	I		1000	1000	1000	1000	1000	5000
	II	0		967	10	1000	1000	2977
	III	0	33		32	567	146	778

	IV	0	990	968		1000	1000	3958
	V	0	0	433	0		124	557
	VI	0	0	854	0	876		1730
	Total	0	14608	27144	6068	23441	17739	

Nos seis diferentes desvios padrões obtemos que o método I possui sempre erro quadrático médio menor do que todos os métodos. Depois deste os métodos que se destacam são, respectivamente, IV e II. O método III possui maior soma das colunas, tendo seu EQMT(k), na maioria das vezes, maior que dos outros métodos. O método V tem a segunda maior soma; no caso do desvio padrão 0.89, este sempre tem EQMT(k) maior que qualquer outro método.

O outro método que segue a ordem de maior EQMT é VI.

Quadro 4.4 Número de vezes em que EQMT(k) de cada um dos métodos da linha é menor que os da coluna, com coeficiente correlação teórico 0.9

σ		I	II	III	IV	V	VI	Total
	I		1000	1000	1000	1000	1000	5000
	II	0		1000	0	986	999	2985
0.55	**III**	0	0		0	94	0	94
	IV	0	1000	1000		999	1000	3999
	V	0	14	906	1		639	1560
	VI	0	1	1000	0	361		1362
	I		1000	1000	1000	1000	1000	5000
	II	0		1000	2	965	1000	2967
0.7	**III**	0	0		0	121	8	129
	IV	0	998	1000		1000	1000	3998
	V	0	35	879	0		791	1705
	VI	0	0	992	0	209		1201
	I		1000	1000	1000	1000	1000	5000
	II	0		1000	8	990	1000	2998
0.89	**III**	0	0		0	435	0	435
	IV	0	992	1000		998	1000	3990
	V	0	10	565	2		357	934
	VI	0	0	1000	0	643		1643
	I		1000	1000	1000	1000	1000	5000
	II	0		943	0	995	1000	2938
1.0	**III**	0	57		28	163	312	560
	IV	0	1000	972		1000	1000	3972
	V	0	5	837	0		1000	1842
	VI	0	0	688	0	0		688

	I		1000	1000	1000	1000	1000	5000
	II	0		1000	2	972	998	2972
3.0	**III**	0	0		0	145	1	146
	IV	0	998	1000		999	1000	3997
	V	0	28	855	1		765	1649
	VI	0	2	999	0	235		1236
	I		1000	1000	1000	1000	1000	5000
	II	0		1000	0	991	1000	2991
5.0	**III**	0	0		0	3	9	12
	IV	0	1000	1000		1000	1000	4000
	V	0	9	997	0		996	2002
	VI	0	0	991	0	4		995
	Total	0	12149	28624	6044	20308	22875	

Como no quadro 4.3, nesse caso de correlação teórica 0.9 também nos fornece o método I com menor EQMT(k). Em seguida os métodos IV e II é que se destacam. Os métodos com maior erro quadrático médio total são: III, VI e V.

Quadro 4.5 Número de vezes em que EQMT(k) de cada um dos métodos da linha é menor que os da coluna, com coeficiente de correlação teórico 0.95

σ		**I**	**II**	**III**	**IV**	**V**	**VI**	**Total**
	I		1000	1000	1000	1000	1000	5000
	II	0		997	0	967	856	2820
0.55	**III**	0	3		0	180	0	183
	IV	0	1000	1000		999	1000	3999
	V	0	33	820	1		478	1332
	VI	0	144	1000	0	522		1666
	I		1000	1000	1000	1000	1000	5000
	II	0		988	2	955	699	2644
0.7	**III**	0	12		0	313	0	325
	IV	0	998	1000		999	1000	3997
	V	0	45	687	1		495	1228
	VI	0	301	1000	0	505		1806
	I		1000	1000	1000	1000	1000	5000
	II	0		1000	0	996	1000	2996
0.89	**III**	0	0		0	12	25	37
	IV	0	1000	1000		1000	1000	4000
	V	0	4	988	0		1000	1992
	VI	0	0	975	0	0		975

	I		1000	1000	1000	1000	1000	5000
	II	0		983	8	827	595	2413
1.0	**III**	0	17		0	402	0	419
	IV	0	992	1000		994	1000	3986
	V	0	173	598	6		430	1207
	VI	0	405	1000	0	570		1975
	I		1000	1000	1000	1000	1000	5000
	II	0		998	2	933	920	2853
3.0	**III**	0	2		0	105	0	107
	IV	0	998	1000		999	1000	3997
	V	0	67	895	1		802	1765
	VI	0	80	1000	0	198		1278
	I		1000	1000	1000	1000	1000	5000
	II	0		998	3	860	886	2747
5.0	**III**	0	2		0	236	1	239
	IV	0	997	1000		998	1000	3995
	V	0	140	764	2		624	1530
	VI	0	114	999	0	376		1489
	Total	0	13527	28690	6026	20946	20811	

Nesse quadro 4.5 de correlação 0.95 também obtemos como melhor estimador "ridge", no sentido de ter menor EQMT, o método I que nos fornece em toda soma um total de 5000. O segundo e terceiro menor EQMT é atribuído, respectivamente, aos métodos IV e II. Já os de maior EQMT temos III, V e VI.

Quadro 4.6 Número de vezes em que EQMT(k) de cada um dos métodos da linha é menor que os da coluna, com coeficiente de correlação teórico 0.99

σ		I	II	III	IV	V	VI	Total
	I		1000	1000	1000	1000	1000	5000
	II	0		992	130	868	639	2629
0.55	**III**	0	8		91	66	0	165
	IV	0	870	909		933	836	3548
	V	0	132	934	67		472	1605
	VI	0	361	1000	164	528		2053
	I		1000	1000	1000	1000	1000	5000
	II	0		994	136	856	651	2637
0.7	**III**	0	6		92	67	0	165
	IV	0	864	908		905	823	3500
	V	0	144	933	95		485	1657
	VI	0	349	1000	177	515		2041

	I		1000	1000	1000	1000	1000	5000
	II	0		958	187	910	657	2712
0.89	**III**	0	42		136	205	0	383
	IV	0	813	864		922	739	3338
	V	0	90	795	78		621	1584
	VI	0	343	1000	261	379		1983
	I		1000	1000	1000	1000	1000	5000
	II	0		997	219	916	795	2927
1.0	**III**	0	3		112	91	0	206
	IV	0	781	888		921	844	3434
	V	0	84	909	79		262	1334
	VI	0	205	1000	156	738		2099
	I		1000	1000	1000	1000	1000	5000
	II	0		991	146	822	675	2634
3.0	**III**	0	9		106	74	0	189
	IV	0	854	894		917	813	3478
	V	0	178	926	83		515	1702
	VI	0	325	1000	187	485		1997
	I		1000	1000	1000	1000	1000	5000
	II	0		992	145	880	684	2701
5.0	**III**	0	8		93	84	0	185
	IV	0	855	907		929	827	3518
	V	0	120	916	71		482	1589
	VI	0	316	1000	173	518		2007
	Total	0	13760	28707	9184	20529	17820	

O quadro 4.6 trata do último caso de correlação, 0.99. Deste modo, como em todos outros casos, obtivemos os métodos I, IV e II com os menores EQMT's e os métodos III, V e VI maiores EQMT's. Vale lembrar que apesar dos métodos III e V possuírem valores maiores do EQMT que dos outros métodos, estes valores são sempre menores que dos estimadores de mínimos quadrados.

Os quatro quadros anteriores nos mostram quantas vezes um determinado método é menor que os demais. Assim, dos 1000 modelos gerados temos quantas vezes um método foi melhor que o outro, o que nos dá a frequência. Do total desta frequência obtemos a porcentagem total que cada um dos métodos é melhor que os demais. Estas porcentagens estão representadas no quadro 4.7, considerando todas correlações e desvios padrões.

Quadro 4.7 Porcentagem em que método de estimação de k tem menor EQMT(k) que os demais, considerando os desvios padrões e as correlações teóricas

σ \ c^2		**0.8**	**0.9**	**0.95**	**0.99**
	I	100	100	100	100
	II	57.58	59.7	56.4	52.58
0.55	**III**	0.52	1.88	3.66	3.3
	IV	80	79.98	79.98	70.96
	V	36.74	31.2	26.64	32.1
	VI	25.16	27.24	33.32	41.06
	I	100	100	100	100
	II	48.28	59.34	52.88	52.74
	III	8.04	2.58	6.5	3.3
0.7	**IV**	79.92	79.96	79.94	70
	V	24.04	34.1	24.56	33.14
	VI	39.72	24.02	36.12	40.82
	I	100	100	100	100
	II	39.78	59.96	59.92	54.24
	III	20.68	8.7	0.74	7.66
0.89	**IV**	79.9	79.8	80	66.76
	V	0	18.68	39.84	31.68
	VI	59.64	32.86	19.5	39.66
	I	100	100	100	100
	II	53.62	58.76	48.26	58.54
	III	0.28	11.2	8.38	4.12
1.0	**IV**	79.66	79.44	79.72	68.68
	V	46.42	36.84	24.14	26.68
	VI	20.02	13.76	39.5	41.98
	I	100	100	100	100
	II	49.04	59.44	57.06	52.68
3.0	**III**	12.04	2.92	2.14	3.78
	IV	80	79.94	79.94	69.56
	V	12.84	32.98	35.3	34.04
	VI	46.08	24.72	25.56	39.94
	I	100	100	100	100
	II	59.54	59.82	54.94	54.02
5.0	**III**	15.56	0.24	4.78	3.7
	IV	79.16	80	79.9	70.36
	V	11.14	40.04	30.6	31.78
	VI	34.6	19.9	29.78	40.14

Poderemos ver com mais clareza o comportamento entre os métodos e com relação ao estimador de mínimos quadrados. Para isso, consideramos M como o número

médio da relação entre o EQMT do "ridge" sobre o EQMT dos mínimos quadrados, isto é, $M = \sum_{i=1}^{1000} \left(\frac{EQMT_i(k)}{EQMT_i(0)}\right) / 1000$ e traçaremos os seguintes gráficos.

Gráfico 4.1 M como função do desvio padrão, com $c^2 = 0.8$

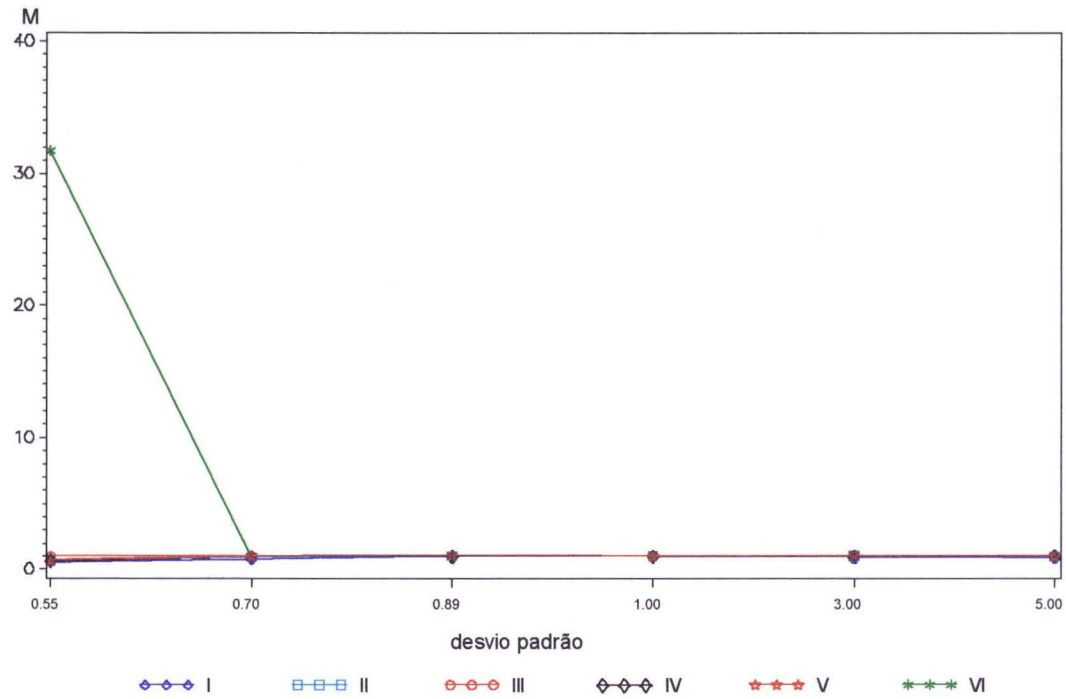
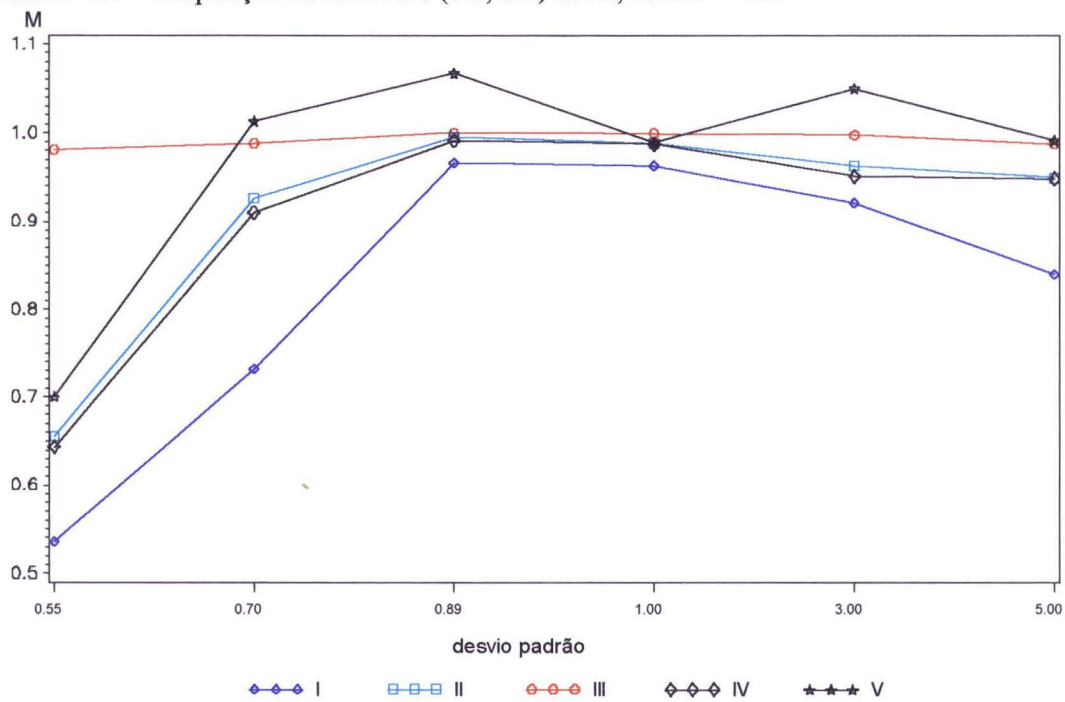
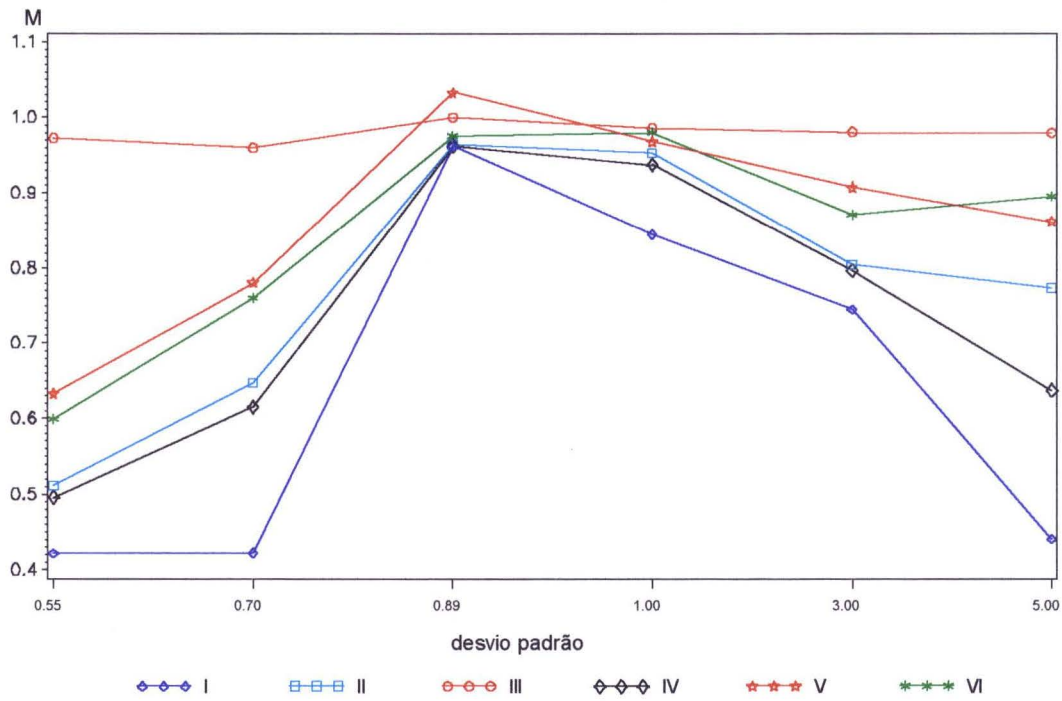


Gráfico 4.2 Ampliaçao do intervalo (0.5, 1.1) de M, com $c^2 = 0.8$



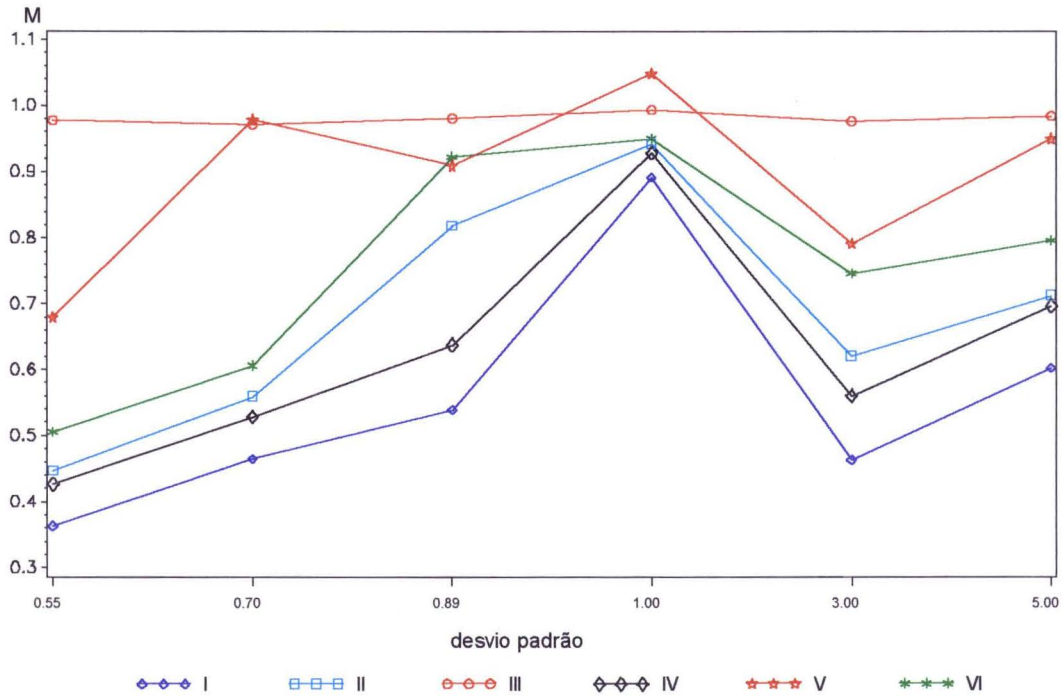
Os dois primeiros gráficos nos mostram o caso da correlação teórica 0.8. No gráfico 4.1 temos a visualização geral incluindo todos os métodos e no gráfico 4.2 mostramos somente a visualização dos valores de M no intervalo (0.5, 1.1). A razão de considerarmos os dois gráficos foi devido ao problema de escala decorrente do método VI possuir o valor de M muito maior que dos outros métodos, quando $\sigma = 0.55$. Assim, temos que somente o método V e VI possuem $M > 1.0$. Os demais métodos possuem os valores de M menores que 1.0, logo, possuem seus EQMT's menores que EQMT(0). Notamos, ainda, que entre os métodos "ridge" , o I possui menor EQMT, neste caso de coeficiente correlação igual a 0.8. O segundo menor EQMT é do método IV. O método III possui seu EQMT, no decorrer de todo desvio padrão, aproximadamente, constante e próximo de 1.0.

Gráfico 4.3 M como função do desvio padrão, com $c^2 = 0.9$



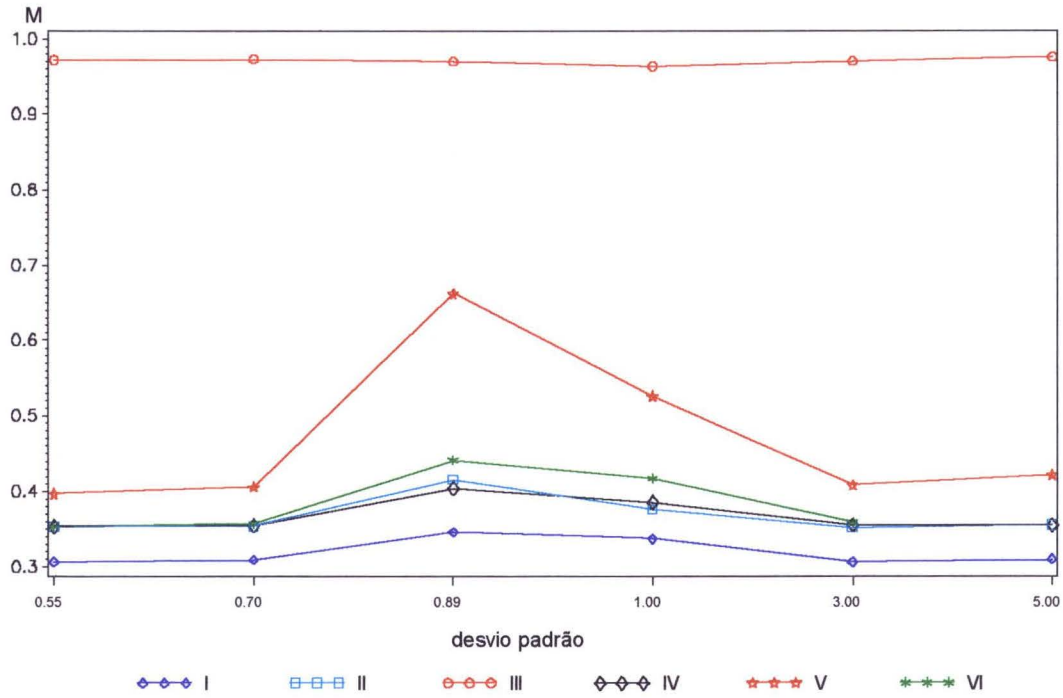
Nesse terceiro gráfico temos a ocorrência do coeficiente correlação igual a 0.9. Novamente o método V obtém valores de $M > 1$, neste caso, quando σ é igual a 0.89. Nos demais métodos temos $M < 1$. Novamente o método I possui menor EQMT e o segundo menor valor de M é do método IV. O método III se apresenta com M, aproximadamente, constante com valor muito próximo de 1.0.

Gráfico 4.4 M como função do desvio padrão, com $c^2 = 0.95$



Neste terceiro gráfico consideramos coeficiente correlação igual a 0.95. O método V possui M > 1, mas agora em $\sigma = 1.0$. Os demais métodos possuem M < 1. Repetem-se as análises feitas quanto ao menor EQMT, confirmando a análise dos quadros.

Gráfico 4.5 M como função do desvio padrão, com $c^2 = 0.99$



Neste quinto gráfico consideramos o caso do coeficiente correlação igual a 0.99. Nele vemos que o método V possui EQMT(k) < EQMT(0) para qualquer valor de σ. Novamente temos que o método III possui M próximo de 1, porém menor.

Analisaremos agora como os métodos se comportam utilizando o EQMTP como medida de comparação. No quadro 4.8 veremos o número de vezes que EQMTP do estimador de mínimos quadrados é menor que do método "ridge".

Quadro 4.8 Número de vezes em que o EQMTP(0) é menor que EQMTP(k)

c^2		I	II	III	IV	V	VI
0.8	**0.55**	0	0	0	0	17	1
	0.7	0	0	0	0	223	0
	0.89	0	0	0	0	1000	0
	1.0	0	0	0	0	6	0
	3.0	0	0	0	0	431	0
	5.0	0	0	0	0	867	0
0.9	**0.55**	0	0	0	1	66	0
	0.7	0	0	0	0	107	0
	0.89	0	0	0	0	425	0
	1.0	0	0	0	0	0	0
	3.0	0	0	0	0	130	0

	5.0	0	0	0	69	0	0
0.95	**0.55**	0	0	0	1	128	0
	0.7	0	0	0	0	233	2
	0.89	0	0	0	8	0	0
	1.0	0	0	0	0	296	0
	3.0	0	0	0	16	85	0
	5.0	0	0	0	0	243	0
0.99	**0.55**	0	0	0	19	45	0
	0.7	0	0	0	34	56	0
	0.89	0	9	0	34	159	0
	1.0	0	0	0	38	367	0
	3.0	0	0	0	25	68	0
	5.0	0	0	0	26	104	0

Neste quadro 4.8 vemos que as colunas dos métodos I, II, III, IV e VI quase sempre apresenta o valor zero. Esse valor corresponde a dizer que EQMTP(k) nunca é maior que EQMTP(0). *Quase sempre* se atribui ao método IV que quando $c^2 = 0.99$ possui uma frequência não nula de EQMTP(0)<EQMTP(k).

Já o método V, como no caso do EQMT, apresenta uma frequência maior que zero de EQMTP(0)<EQMTP(k). Entretanto, esta frequência é maior quando a correlação teórica entre as regressoras é 0.8 e menor quando $c^2 = 0.99$.

Novamente, agora com a variável EQMTP, obtemos este valor dos estimadores "ridge" menor que dos mínimos quadrados, independentemente, do mal condicionamento da matriz X.

Para uma melhor análise dessa variável, faremos a comparação entre os métodos "ridge", isto é, veremos a frequência das vezes que o EQMTP, por exemplo, do método I é menor que cada um dos outros métodos.

Quadro 4.9 Número de vezes em que EQMTP(k) de cada um dos métodos da linha é menor que os da coluna, com coeficiente de correlação teórico 0.8.

σ		**I**	**II**	**III**	**IV**	**V**	**VI**	**Total**
	I		1000	1000	1000	1000	1000	5000
	II	0		1000	717	913	999	3629
0.55	**III**	0	0		0	20	1	21
	IV	0	283	1000		732	997	3012
	V	0	87	980	268		859	2194
	VI	0	1	999	3	141		1144

	I		1000	1000	1000	1000	1000	5000
	II	0		994	353	939	964	3250
0.7	**III**	0	6		3	250	14	273
	IV	0	647	997		847	1000	3491
	V	0	61	750	153		592	1556
	VI	0	36	986	0	408		1430
	I		1000	1000	1000	1000	1000	5000
	II	0		830	0	1000	1000	2830
0.89	**III**	0	170		135	1000	188	1493
	IV	0	1000	865		1000	1000	3865
	V	0	0	0	0		0	0
	VI	0	0	812	0	1000		1812
	I		1000	1000	1000	1000	1000	5000
	II	0		998	1000	993	1000	3991
1.0	**III**	0	2		2	15	5	24
	IV	0	0	998		942	1000	2940
	V	0	7	985	58		734	1784
	VI	0	0	995	0	266		1261
	I		1000	1000	1000	1000	1000	5000
	II	0		999	97	973	966	3035
3.0	**III**	0	1		0	442	0	443
	IV	0	903	1000		955	1000	3858
	V	0	27	558	45		327	957
	VI	0	34	1000	0	673		1707
	I		1000	1000	1000	1000	1000	5000
	II	0		936	658	1000	1000	3594
5.0	**III**	0	64		67	924	185	1240
	IV	0	342	933		1000	1000	3275
	V	0	0	76	0		18	94
	VI	0	0	815	0	982		1797
	Total	0	9671	26506	9559	23415	20849	

Observamos no total das linhas que os maiores valores correspondem aos métodos que possuem maior número de vezes seus EQMTP's menores que dos outros métodos e os menores valores são que os possuem os maiores EQMTP's, isto para cada um dos desvios. Os métodos que se atribuem ao primeiro caso é I e no segundo caso são III e V. A soma das colunas mostram-nos os métodos que se apresentam com menor e maior valor, agora no caso da correlação 0.8. Os métodos que se encaixam nesta situação são, respectivamente, I e III. Logo, I possui menor valor do EQMTP e III o maior valor.

Quadro 4.10 Número de vezes em que EQMTP(k) de cada um dos métodos da linha é menor que os da coluna, com coeficiente de correlação teórico 0.9.

σ		I	II	III	IV	V	VI	Total
	I		1000	1000	1000	1000	1000	5000
	II	0		1000	460	858	1000	3318
0.55	**III**	0	0		1	70	0	71
	IV	0	540	999		802	991	3332
	V	0	142	930	198		793	2063
	VI	0	0	1000	9	207		1216
	I		1000	1000	1000	1000	1000	5000
	II	0		1000	312	965	996	3273
0.7	**III**	0	0		1	127	7	135
	IV	0	688	999		894	968	3549
	V	0	35	873	106		802	1816
	VI	0	4	993	32	198		1227
	I		1000	1000	1000	1000	1000	5000
	II	0		1000	33	995	1000	3028
0.89	**III**	0	0		0	425	0	425
	IV	0	967	1000		990	1000	3957
	V	0	5	575	10		374	964
	VI	0	0	1000	0	626		1626
	I		1000	1000	1000	1000	1000	5000
	II	0		946	10	996	1000	2952
1.0	**III**	0	54		30	142	281	507
	IV	0	990	970		995	1000	3955
	V	0	4	858	5		1000	1867
	VI	0	0	719	0	0		719
	I		1000	1000	1000	1000	1000	5000
	II	0		1000	106	994	1000	3100
3.0	**III**	0	0		0	142	1	143
	IV	0	894	1000		973	1000	3867
	V	0	6	858	27		766	1657
	VI	0	0	999	0	234		1233
	I		1000	1000	1000	1000	1000	5000
	II	0		1000	238	997	1000	3235
5.0	**III**	0	0		71	4	7	82
	IV	0	762	929		871	889	3451
	V	0	3	996	129		996	2124
	VI	0	0	993	111	4		1108
	Total	0	11094	28637	7889	19509	22871	

Para este caso onde $c^2 = 0.9$, os resultados se repetem aos obtidos quando $c^2 = 0.8$. Desta forma, os de menor EQMTP é o método I e os de maior EQMTP é III. Observamos também que depois do método I, IV possui EQMTP menor que dos outros métodos.

Quadro 4.11 Número de vezes em que EQMTP(k) de cada um dos métodos da linha é menor que os da coluna, com coeficiente de correlação teórico 0.95.

σ		I	II	III	IV	V	VI	Total
	I		1000	1000	1000	1000	1000	5000
	II	0		1000	564	790	975	3329
0.55	III	0	0		1	133	0	134
	IV	0	436	999		688	955	3078
	V	0	210	867	312		669	2058
	VI	0	25	1000	45	331		1401
	I		1000	1000	1000	1000	1000	5000
	II	0		999	485	833	902	3219
0.7	III	0	1		0	248	2	251
	IV	0	515	1000		729	937	3181
	V	0	167	752	271		612	1802
	VI	0	98	998	63	388		1547
	I		1000	1000	1000	1000	1000	5000
	II	0		1000	54	1000	1000	3054
0.89	III	0	0		10	16	31	57
	IV	0	946	990		982	983	3901
	V	0	0	984	18		1000	2002
	VI	0	0	969	17	0		986
	I		1000	1000	1000	1000	1000	5000
	II	0		996	235	916	896	3043
1.0	III	0	4		0	313	3	320
	IV	0	765	1000		866	1000	3631
	V	0	84	687	134		548	1453
	VI	0	104	997	0	452		1553
	I		1000	1000	1000	1000	1000	5000
	II	0		1000	360	972	986	3318
3.0	III	0	0		16	89	0	105
	IV	0	640	984		878	971	3473
	V	0	28	911	122		827	1888
	VI	0	14	1000	29	173		1216
	I		1000	1000	1000	1000	1000	5000
	II	0		999	412	971	972	3354

5.0	**III**	0	1		0	254	1	256
	IV	0	588	1000		823	1000	3411
	V	0	29	746	177		586	1538
	VI	0	28	999	0	414		1441
	Total	0	10683	28877	9325	19259	21856	

Novamente obtemos os métodos I e IV como os de menor EQMTP e o método III com maior EQMTP.

Quadro 4.12 Número de vezes em que EQMTP(k) de cada um dos métodos da linha é menor que os da coluna, com coeficiente de correlação teórico 0.99.

σ		I	II	III	IV	V	VI	Total
	I		1000	1000	1000	1000	1000	5000
	II	0		996	643	534	824	2997
0.55	**III**	0	4		91	51	1	147
	IV	0	357	909		363	776	2405
	V	0	466	949	637		720	2772
	VI	0	176	999	224	280		1679
	I		1000	1000	1000	1000	1000	5000
	II	0		998	664	627	836	3125
0.7	**III**	0	2		92	65	0	159
	IV	0	336	908		379	753	2376
	V	0	373	935	621		699	2628
	VI	0	164	1000	247	301		1712
	I		1000	1000	1000	1000	1000	5000
	II	0		983	682	806	856	3327
0.89	**III**	0	17		137	162	0	316
	IV	0	318	863		677	715	2573
	V	0	194	838	323		748	2103
	VI	0	144	1000	285	252		1681
	I		1000	1000	1000	1000	1000	5000
	II	0		1000	540	918	692	3150
1.0	**III**	0	0		112	388	0	500
	IV	0	460	888		752	619	2719
	V	0	82	612	248		288	1230
	VI	0	308	1000	381	712		2401
	I		1000	1000	1000	1000	1000	5000
	II	0		999	677	687	843	3206
3.0	**III**	0	1		106	74	0	181

	IV	0	323	894		444	744	2405
	V	0	313	926	556		718	2513
	VI	0	157	1000	256	282		1695
	I		1000	1000	1000	1000	1000	5000
	II	0		1000	653	614	803	3070
5.0	**III**	0	0		93	109	0	202
	IV	0	347	907		459	722	2435
	V	0	386	891	541		644	2462
	VI	0	197	1000	278	356		1831
	Total	0	11125	28495	15087	16292	19001	

Para o caso de correlação 0.99 os resultados permanecem como nos anteriores. Logo, o método I continua com menor EQMTP. Em segundo lugar, nesta ocorrência, está o método II, o de maior EQMTP foi atribuído ao método III.

Os resultados destes quadros podem ser resumidos no quadro 4.13, onde consideramos todos métodos e suas respectivas porcentagens, como no quadro 4.7.

Quadro 4.13 Porcentagem de vezes em que o EQMTP de um método é menor que dos outros, considerando os desvios padrões e as correlações teóricas.

σ \ c^2		**0.8**	**0.9**	**0.95**	**0.99**
	I	100	100	100	100
	II	72.58	66.36	66.58	59.94
0.55	**III**	0.42	1.42	2.68	2.94
	IV	60.24	66.64	61.56	48.1
	V	43.88	41.26	41.16	55.44
	VI	22.88	24.32	28.02	33.58
	I	100	100	100	100
	II	65	65.46	64.38	62.5
	III	5.46	2.7	5.02	3.18
0.7	**IV**	69.82	70.98	63.62	47.52
	V	31.12	36.32	36.04	52.56
	VI	28.6	24.54	30.94	34.24
	I	100	100	100	100
	II	56.6	60.56	61.08	66.54
	III	29.86	8.5	1.14	6.32
0.89	**IV**	77.3	79.14	78.02	51.46
	V	0	19.28	40.04	42.06
	VI	36.24	32.52	19.72	33.62
	I	100	100	100	100

	II	79.82	59.04	60.86	63
	III	0.48	10.14	6.4	10
1.0	IV	58.8	79.1	72.62	54.38
	V	35.68	37.34	29.06	24.6
	VI	25.22	14.38	31.06	48.02
	I	100	100	100	100
	II	60.7	62	66.36	64.12
3.0	III	8.86	2.86	2.1	3.62
	IV	77.16	77.34	69.46	48.1
	V	19.14	33.14	37.76	50.26
	VI	34.14	24.66	24.32	33.9
	I	100	100	100	100
	II	71.88	64.7	67.08	61.4
5.0	III	24.8	1.64	5.12	4.04
	IV	65.5	69.02	68.22	48.7
	V	1.88	42.48	30.76	49.24
	VI	35.94	22.16	28.82	36.62

Podemos ver graficamente o desempenho entre os métodos e com relação ao estimador de mínimos quadrados. Considerando M como o número médio da relação entre o EQMTP do "ridge" sobre o EQMTP dos mínimos quadrados, isto é, $M = \sum_{i=1}^{1000} \left(\frac{EQMTP_i(k)}{EQMTP_i(0)}\right) / 1000$, traçaremos os seguintes gráficos.

Gráfico 4.6 M como função do desvio padrão com $c^2 = 0.8$

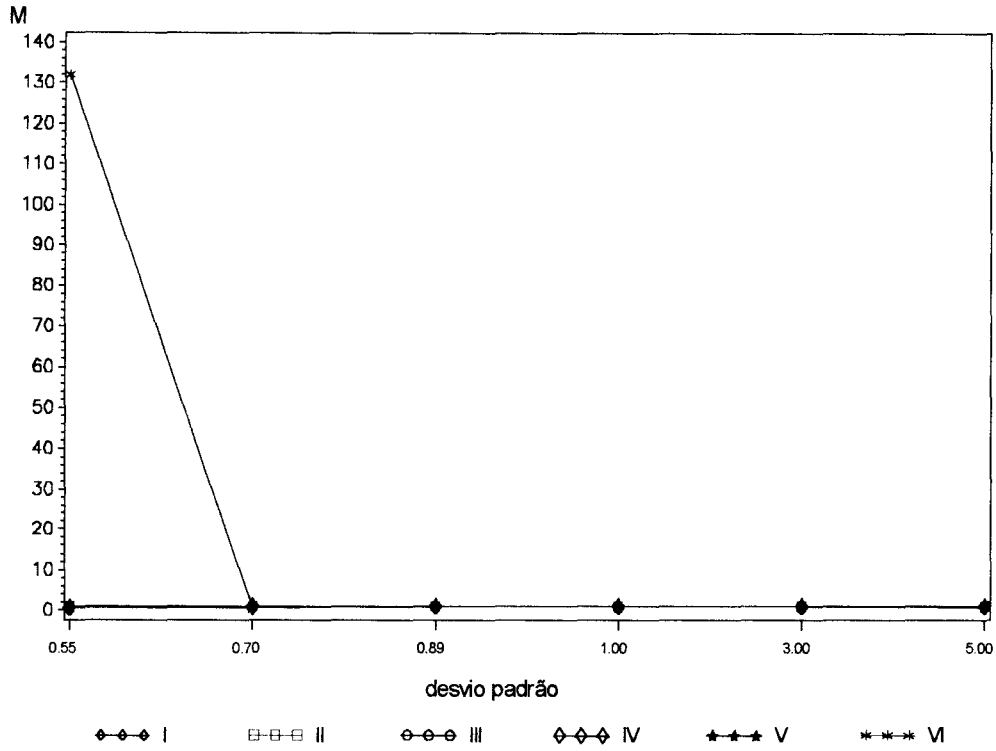
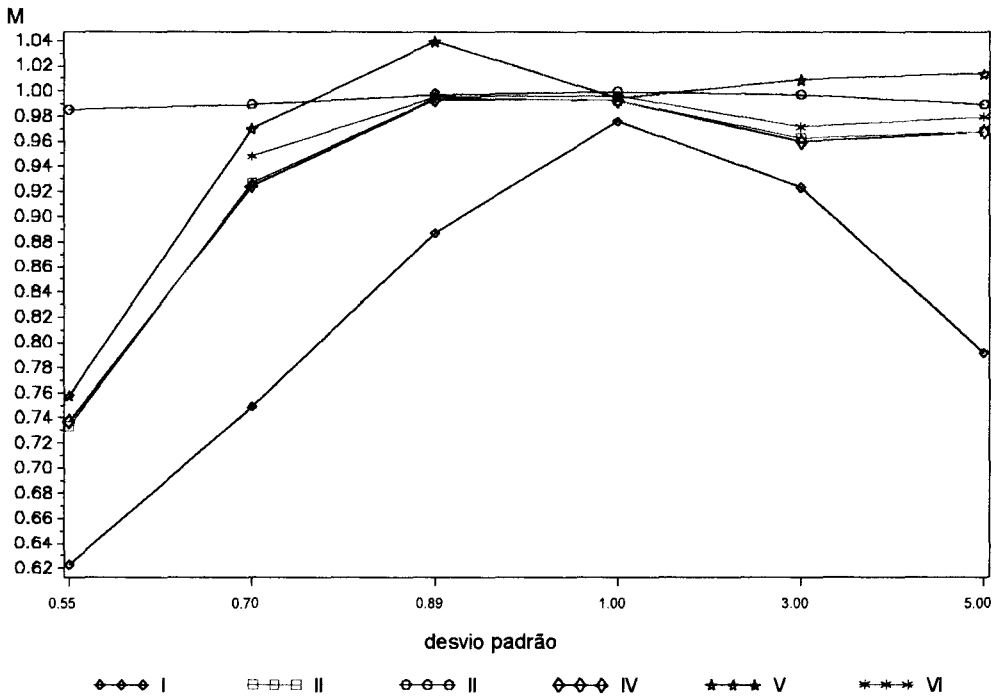


Gráfico 4.7 Ampliação de M no intervalo (0.62, 1.04) com $c^2 = 0.8$



Observando a semelhança dos resultados em todos os gráficos optaremos em comentá-los no final para evitarmos repetições.

Gráfico 4.8 M como função do desvio padrão com $c^2 = 0.9$

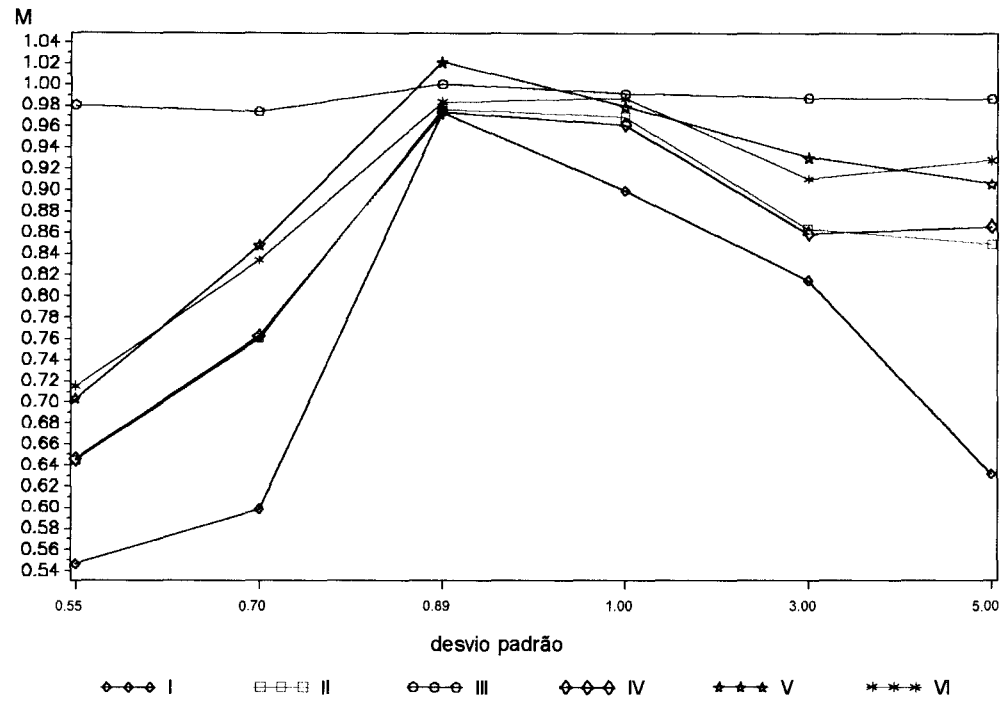


Gráfico 4.9 M como função do desvio padrão com $c^2 = 0.95$

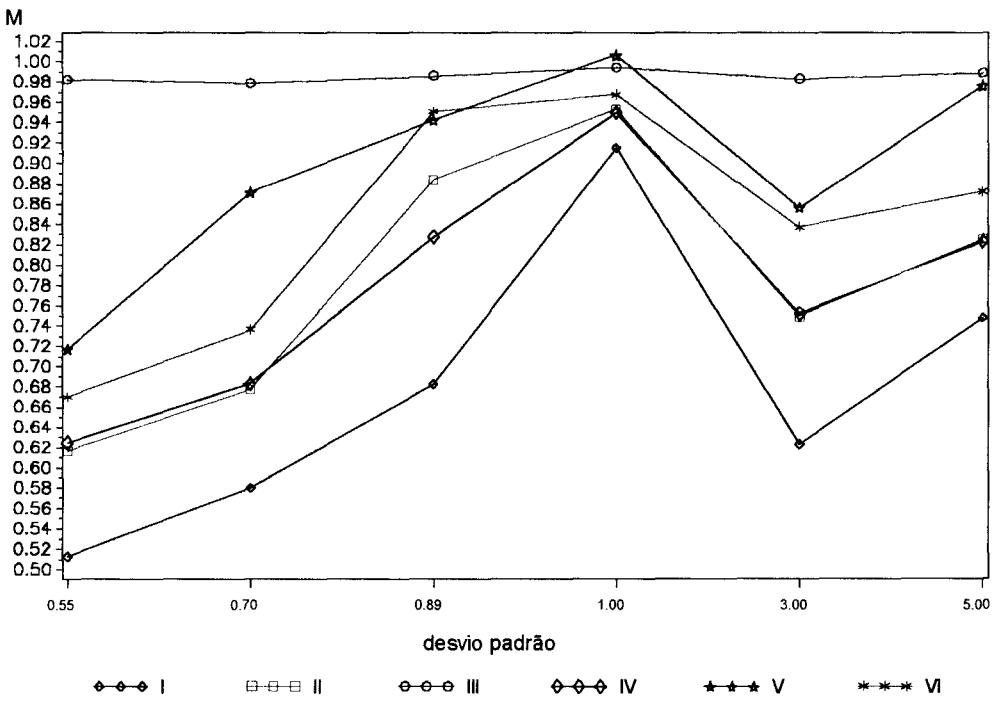
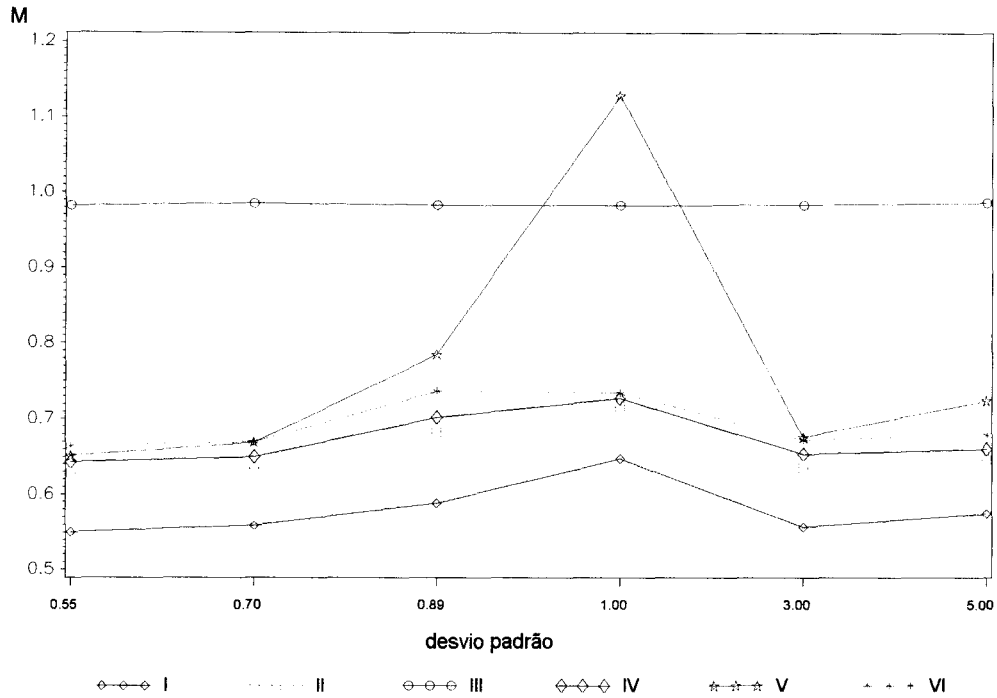


Gráfico 4.10 M como função do desvio padrão com $c^2 = 0.99$



Os gráficos 4.6 a 4.10 mostram o comportamento da variável EQMTP em todos os casos do desvio padrão. Observe que traçamos para o caso da correlação teórica 0.8, dois gráficos. O primeiro nos mostra a visualização geral de todos métodos e o segundo a ampliação do intervalo (0.62, 1.04) de M, pois no primeiro não conseguimos visualizar as definições dos métodos. Isto ocorreu porque quando $\sigma = 0.55$ o método VI possui um valor muito maior comparado aos outros métodos , provocando o problema de visualização causado pela escala. Assim, para retificar esta deficiência ampliamos separadamente no segundo gráfico o comportamento destes sem a observação do método VI, quando $\sigma=0.55$. Desta forma poderemos comparar os valores de M dos diferentes métodos.

Em todos os gráfico vemos que, em média, I possui o menor valor do EQMTP. Em segundo e terceiro lugares os métodos II e IV possuem menor valor. O método V em todas correlações possui desvios com valores de M maior que 1.0. Já o método III apesar de ter $M < 1$, possui quase sempre EQMTP(k) maior que dos outros métodos.

Agora, para uma melhor análise dos estimadores, faremos a decomposição do EQMT(k). Analisaremos o quanto cada método está sendo viciado e qual sua variância. Desse modo, nos próximos quadros veremos a análise do vício.

Devido a semelhança dos resultados, faremos os comentários somente no final dos quadros.

Quadro 4.14 Número de vezes em que Vício de cada um dos métodos da linha é menor que os da coluna, com coeficiente de correlação 0.8.

σ		I	II	III	IV	V	VI	Total
	I		574	0	623	586	25	1808
	II	426		0	573	543	2	1544
0.55	III	1000	1000		1000	1000	1000	5000
	IV	377	427	0		511	2	1317
	V	414	457	0	489		1	1361
	VI	975	998	0	998	999		3970
	I		761	0	265	768	0	1794
	II	239		0	100	719	0	1058
0.7	III	1000	1000		1000	1000	1000	5000
	IV	735	900	0		827	0	2462
	V	232	281	0	173		0	686
	VI	1000	1000	0	1000	1000		4000
	I		676	0	2	1000	0	1678
	II	324		0	0	1000	0	1324
0.89	III	1000	1000		1000	1000	1000	5000
	IV	998	1000	0		1000	0	2998
	V	0	0	0	0		0	0
	VI	1000	1000	0	1000	1000		4000
	I		0	0	0	70	0	70
	II	1000		0	1000	1000	0	3000
1.0	III	1000	1000		1000	1000	999	4999
	IV	1000	0	0		23	0	1023
	V	930	0	0	977		0	1907
	VI	1000	1000	1	1000	1000		4001
	I		955	0	3	953	0	1911
	II	45		0	11	942	0	998
3.0	III	1000	1000		1000	1000	1000	5000
	IV	997	989	0		976	0	2962
	V	47	58	0	24		0	129
	VI	1000	1000	0	1000	1000		4000
	I		23	0	0	827	0	850
	II	977		0	658	1000	0	2635
5.0	III	1000	1000		1000	1000	998	4998
	IV	1000	342	0		1000	0	2342
	V	173	0	0	0		0	173

	VI	1000	1000	12	1000	1000		4012
	Total	21889	19441	13	16824	25674	6027	

Quadro 4.15 Número de vezes em que Vício de cada um dos métodos da linha é menor que os da coluna, com coeficiente de correlação 0.9.

σ		**I**	**II**	**III**	**IV**	**V**	**VI**	**Total**
	I		717	0	865	657	23	2262
	II	283		0	418	594	0	1295
0.55	**III**	1000	1000		1000	1000	1000	5000
	IV	135	582	0		589	0	1306
	V	343	406	0	411		0	1160
	VI	977	1000	0	1000	1000		3977
	I		634	0	887	514	7	2042
	II	366		0	667	395	0	1428
0.7	**III**	1000	1000		1000	1000	1000	5000
	IV	113	333	0		374	0	820
	V	486	605	0	626		3	1720
	VI	993	1000	0	1000	997		3990
	I		927	0	0	927	0	1854
	II	73		0	48	927	0	1048
0.89	**III**	1000	1000		1000	1000	1000	5000
	IV	1000	952	0		935	0	2887
	V	73	73	0	65		0	211
	VI	1000	1000	0	1000	1000		4000
	I		0	0	0	0	0	0
	II	1000		0	1000	8	0	2008
1.0	**III**	1000	1000		1000	965	884	4849
	IV	1000	0	0		3	0	1003
	V	1000	992	35	997		0	3024
	VI	1000	1000	116	1000	1000		4116
	I		394	0	147	436	0	977
	II	606		0	550	462	0	1618
3.0	**III**	1000	1000		1000	1000	1000	5000
	IV	853	450	0		456	0	1759
	V	564	538	0	544		0	1646
	VI	1000	1000	0	1000	1000		4000
	I		39	0	821	16	0	876
	II	961		0	993	18	0	1972
5.0	**III**	1000	1000		1000	993	983	4976
	IV	179	7	0		11	0	197
	V	984	982	7	989		0	2962
	VI	1000	1000	17	1000	1000		4017

	Total	21989	20676	175	22028	19277	5900	

Quadro 4.16 Número de vezes em que Vício de cada um dos métodos da linha é menor que os da coluna, com coeficiente de correlação 0.95.

σ		I	II	III	IV	V	VI	Total
	I		836	0	866	783	14	2499
	II	164		0	283	718	0	1165
0.55	**III**	1000	1000		1000	1000	1000	5000
	IV	134	717	0		725	0	1576
	V	217	282	0	275		1	775
	VI	986	1000	0	1000	999		3985
	I		738	0	667	685	18	2108
	II	262		0	349	639	4	1254
0.7	**III**	1000	1000		1000	1000	1000	5000
	IV	333	651	0		652	4	1640
	V	315	361	0	348		14	1038
	VI	982	996	0	996	986		3960
	I		39	0	578	1	0	618
	II	961		0	990	0	0	1951
0.89	**III**	1000	1000		1000	973	950	4923
	IV	422	10	0		1	0	433
	V	999	1000	27	999		0	3025
	VI	1000	1000	50	1000	1000		4050
	I		844	0	5	742	0	1591
	II	156		0	89	686	0	931
1.0	**III**	1000	1000		1000	1000	1000	5000
	IV	995	911	0		815	0	2721
	V	258	314	0	185		0	757
	VI	1000	1000	0	1000	1000		4000
	I		484	0	792	368	0	1644
	II	516		0	624	298	0	1438
3.0	**III**	1000	1000		1000	999	999	4998
	IV	208	376	0		342	0	926
	V	632	702	1	658		0	1993
	VI	1000	1000	1	1000	1000		4001
	I		676	0	451	715	0	1842
	II	324		0	482	720	0	1526
5.0	**III**	1000	1000		1000	1000	1000	5000
	IV	549	518	0		631	0	1698
	V	285	280	0	369		0	934
	VI	1000	1000	0	1000	1000		4000
	Total	19698	10962	79	21006	21478	6004	

Quadro 4.17 Número de vezes em que Vício de cada um dos métodos da linha é menor que os da coluna, com coeficiente de correlação 0.99

σ		I	II	III	IV	V	VI	Total
	I		943	0	504	941	1	2389
	II	57		0	216	923	0	1196
0.55	**III**	1000	1000		909	1000	1000	4909
	IV	496	784	91		879	164	2414
	V	59	77	0	121		3	260
	VI	999	1000	0	836	997		3832
	I		882	0	504	844	0	2230
	II	118		0	258	748	0	1124
0.7	**III**	1000	1000		908	1000	1000	4908
	IV	496	742	92		807	177	2314
	V	156	252	0	193		0	601
	VI	1000	1000	0	823	1000		3823
	I		593	0	443	500	0	1536
	II	407		0	456	457	0	1320
0.89	**III**	1000	1000		864	1000	1000	4864
	IV	557	544	136		516	261	2014
	V	500	543	0	484		0	1527
	VI	1000	1000	0	739	1000		3739
	I		891	0	442	938	0	2271
	II	109		0	181	856	0	1146
1.0	**III**	1000	1000		888	1000	1000	4888
	IV	558	819	112		930	156	2575
	V	62	144	0	70		0	276
	VI	1000	1000	0	844	1000		3844
	I		837	0	528	786	1	2152
	II	163		0	277	631	0	1071
3.0	**III**	1000	1000		894	1000	1000	4894
	IV	472	723	106		741	187	2229
	V	214	369	0	259		0	842
	VI	999	1000	0	813	1000		3812
	I		922	0	499	943	0	2364
	II	78		0	214	924	0	1216
5.0	**III**	1000	1000		907	1000	1000	4907
	IV	501	786	93		890	173	2443
	V	57	76	0	110		1	244
	VI	1000	1000	0	827	999		3826
	Total	17058	22997	630	16051	26250	11124	

Nestes quadros vemos que o método III obtém sempre vício menor que todos outros métodos. Visualizando os resultados obtidos pelo EQMT e EQMTP, observamos que este método possui, na maioria das vezes, seus valores muito grandes, comparados aos outros métodos. Vimos também que ele aparece nos gráficos com o valor de M muito próximo de 1. Agora, voltando ao vício, temos que ele possui o menor valor, logo estas observações o fazem um método que mais se assemelha com os estimadores de mínimos quadrados. Este fato se deve ao seu critério de divergência-convergência, que assume em vários casos o valor zero para k_i. Depois deste, o que nos fornece menor vício é o método VI. Na análise vemos também que o maior vício é atribuído ao método V e, quando c^2 = 0.9 ao IV.

No próximo quadro veremos o resumo dos resultados obtidos pelos quadros anteriores. Nele temos as informações das porcentagens que cada método obteve, considerando todos os casos dos desvios e correlações.

Quadro 4.18 Porcentagem das vezes em que o método teve Vício maior que os outros, considerando os desvios padrões e as correlações teóricas.

σ \ c^2		0.8	0.9	0.95	0.99
	I	87.9	76.64	67.62	59.5
	II	62	69.14	73.02	73.44
0.55	III	0.02	0	0	1.82
	IV	61.7	59.36	56.2	47.84
	V	68.34	74.86	83.14	94.06
	VI	20.04	20	20.02	23.34
	I	94.9	92.7	71.8	62.34
	II	64.2	60.14	71.92	75.44
	III	0.06	0.12	0	1.84
0.7	IV	45.46	67.28	57.48	49.92
	V	75.44	59.8	78.36	86.92
	VI	19.94	19.96	20.44	23.54
	I	82.1	63.5	99.7	76.06
	II	63.76	78.62	60.2	71.84
	III	0.1	0	0.72	2.72
0.89	IV	40	42.26	80.08	55.5
	V	94.14	95.62	39.78	68.66
	VI	19.9	20	19.52	25.22
	I	100	100	86.66	64.42
	II	40	59.32	66.12	74.04
	III	0.12	8.76	0	2.24
1.0	IV	61.16	79.94	45.48	42.34
	V	78.84	37.76	81.74	93.84
	VI	19.88	14.22	20	23.12
	I	71.46	89.2	90.1	64.8
	II	72.72	63.62	64.82	76.52
3.0	III	0	0.02	0	2.12
	IV	40.7	61.88	67.34	51.1
	V	95.12	65.3	57.74	81.72
	VI	20	19.98	20	23.74
	I	96.68	99.88	81.18	60.8
	II	46.8	59.82	63.66	73.06
5.0	III	2.58	0.28	0	1.86
	IV	53.12	79.66	57.06	46.68
	V	83.26	40.54	78.1	94.12
	VI	17.56	19.82	20	23.48

Graficamente, veremos com mais clareza o comportamento de todos métodos. Deste modo, nos gráficos 4.14 a 4.21 mostraremos o vício como função do desvio padrão

para cada correlação. Ressaltamos que os valores obtidos do vício para cada correlação e desvio padrão tratam-se do valor médio do vício considerando os 1000 modelos de regressão.

Primeiramente, mostraremos o gráfico do ponto de vista geral incluindo todos desvios padrões. Depois, excluímos os valores de σ maior que 1.0 e fazemos um segundo gráfico. O motivo para tal procedimento decorreu da dificuldade de visualizar o comportamento dos outros métodos, pois quando σ assume os valores maiores que 1.0 o valor médio do vício cresce bruscamente, criando um problema de escala. Este problema é a causa de uma amplitude muito grande.

Gráfico 4.11 Vício como função do desvio padrão com $c^2 = 0.8$

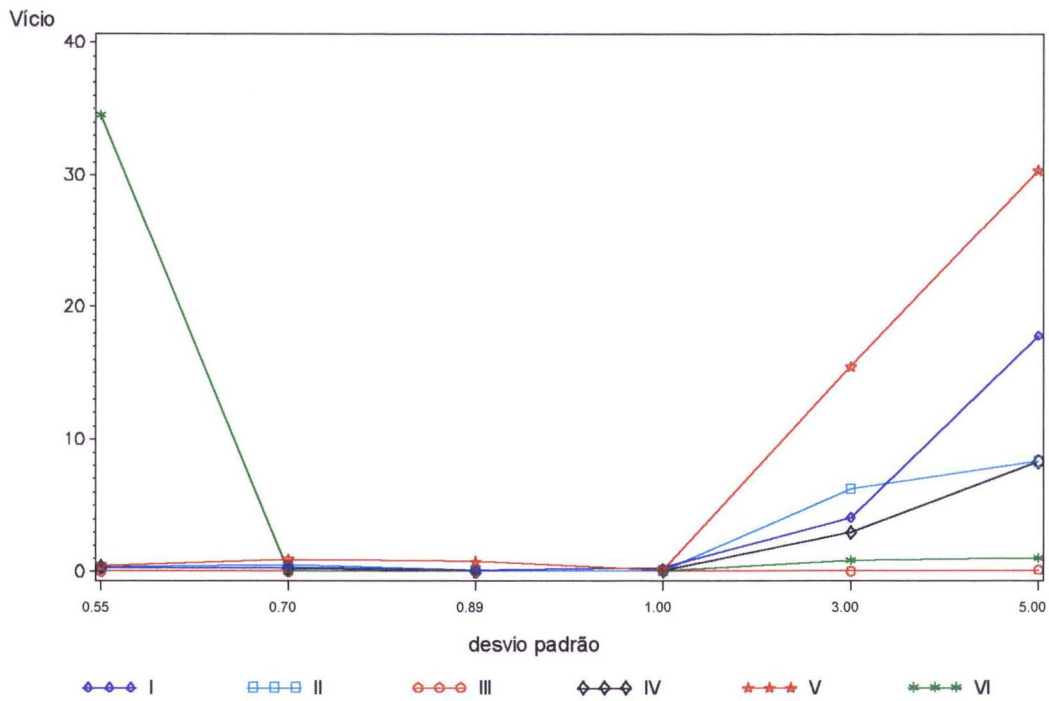


Gráfico 4.12 Ampliação do intervalo (0.55, 1.0) com $c^2 = 0.8$

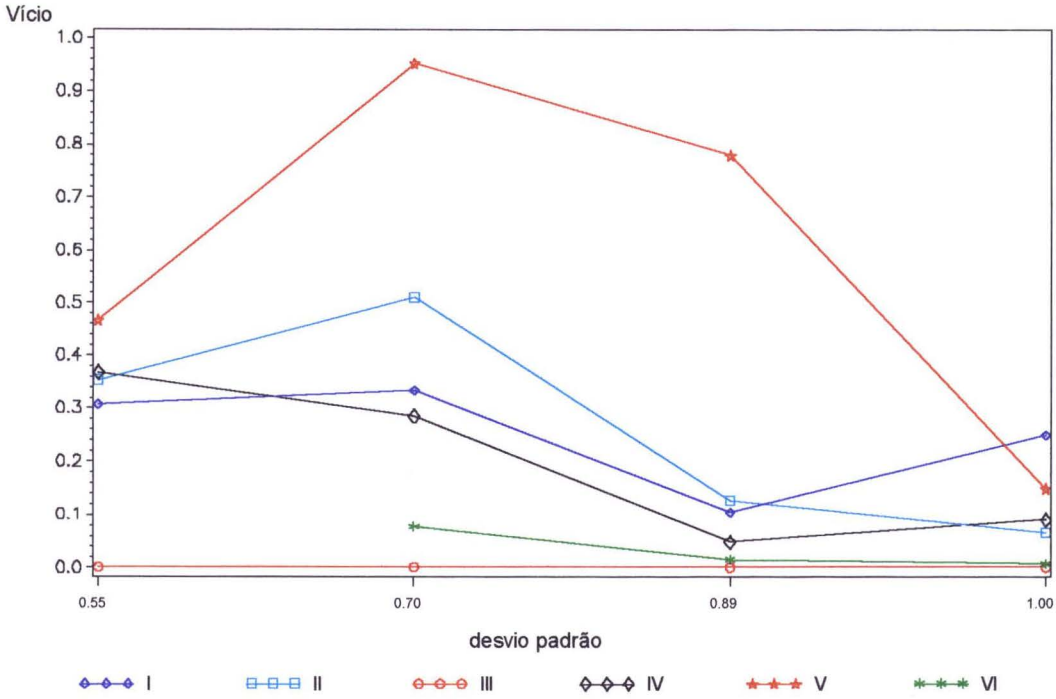


Gráfico 4.13 Vício como função do desvio padrão com $c^2 = 0.9$

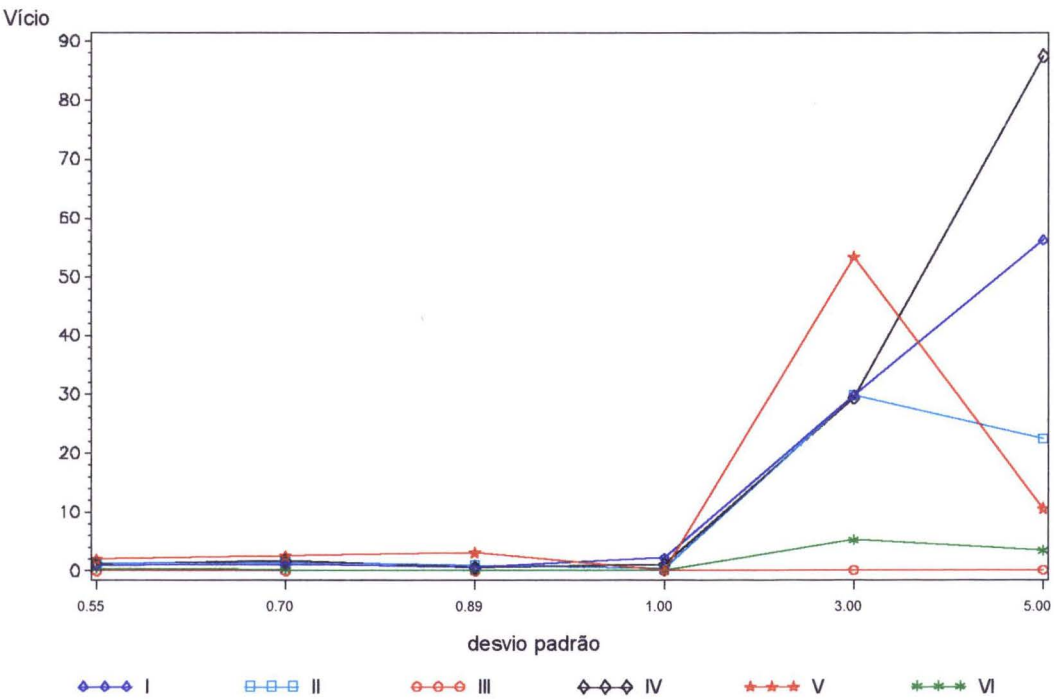


Gráfico 4.14 Ampliação do intervalo (0.55, 1.0) com $c^2 = 0.9$

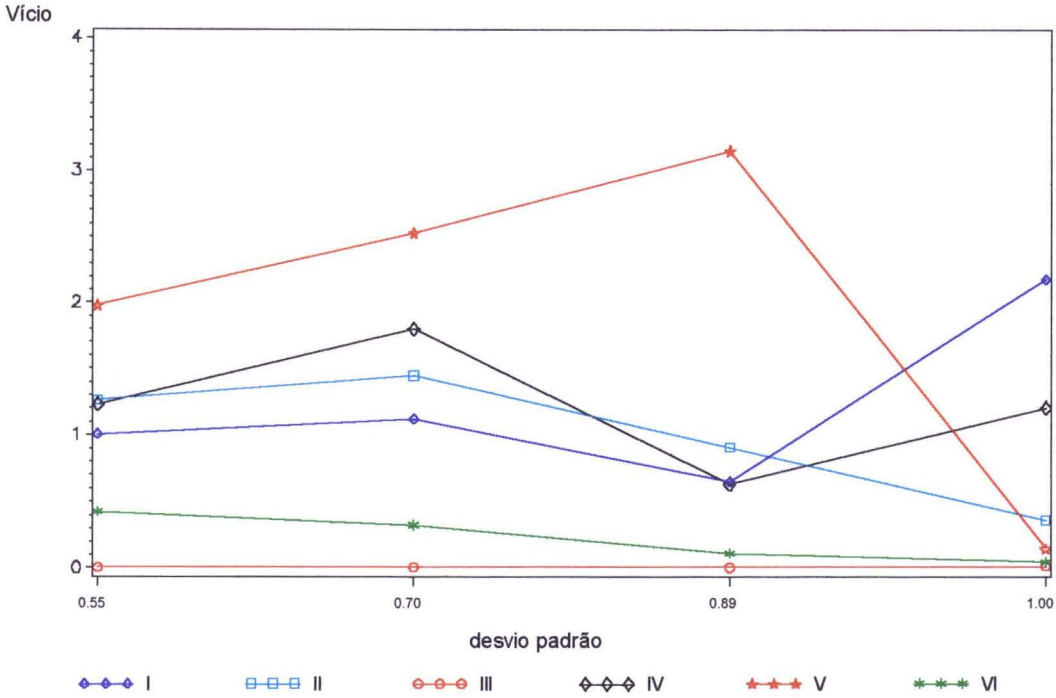


Gráfico 4.15 Vício como função do desvio padrão com $c^2 = 0.95$

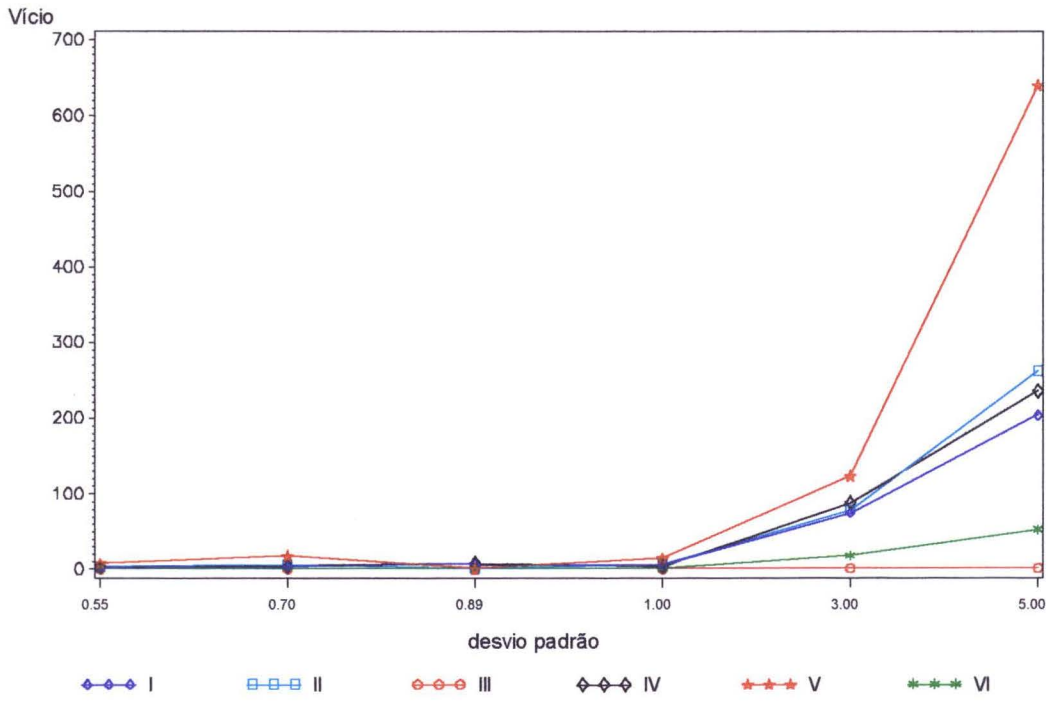


Gráfico 4.16 Ampliação do intervalo (0.55, 1.0) com $c^2 = 0.95$

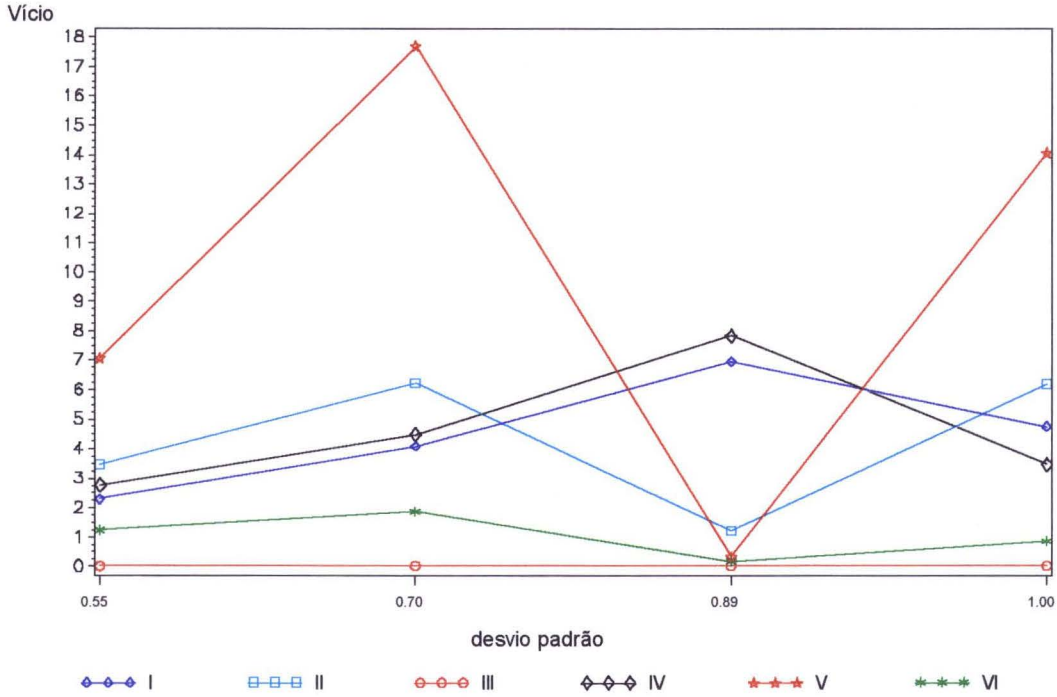


Gráfico 4.17 Vício como função do desvio padrão com $c^2 = 0.99$

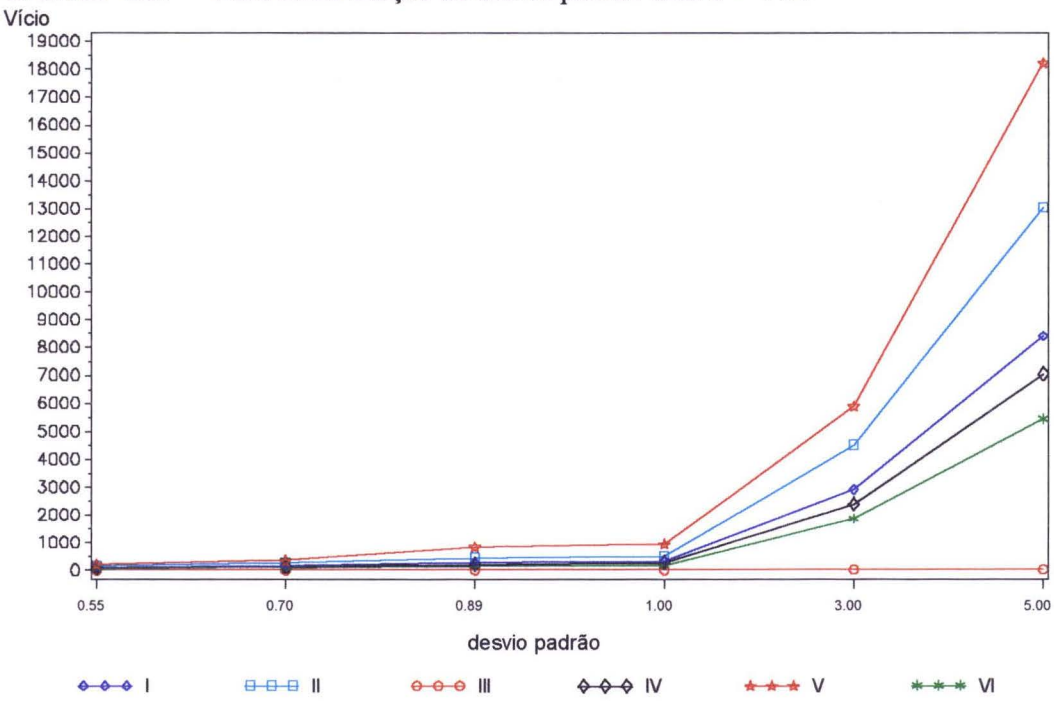
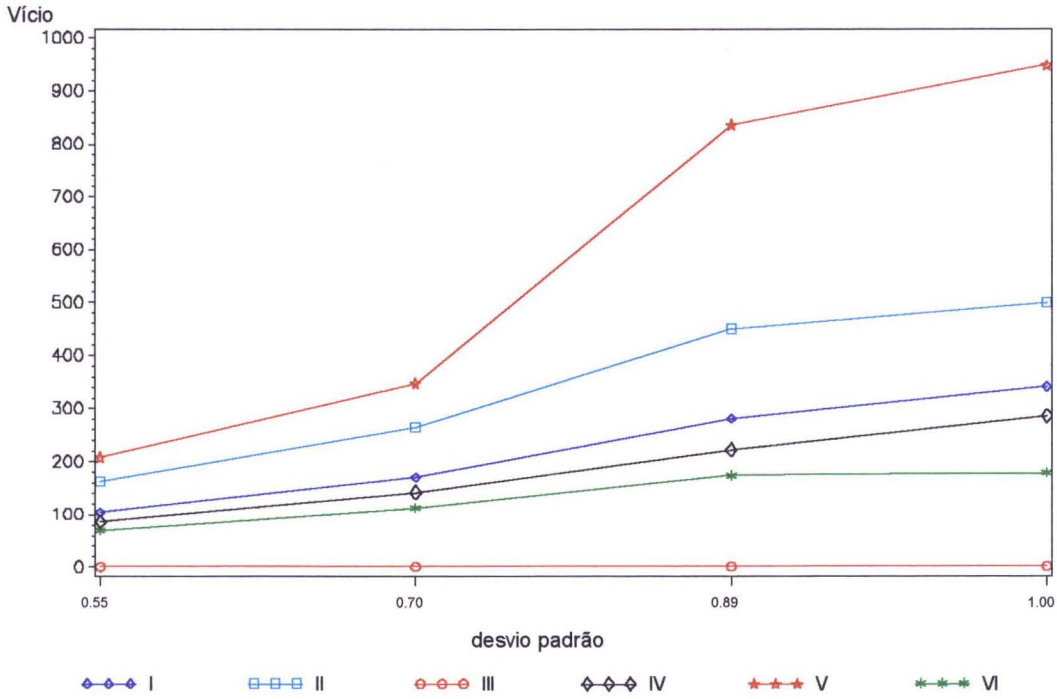


Gráfico 4.18 Ampliação do intervalo (0.55, 1.0) com $c^2 = 0.99$



Os gráficos ilustram os resultados obtidos pelos quadros, isto é, em todos os casos temos que o método III possui o menor vício médio. O segundo menor vício médio é do método VI, com exeção do caso de $\sigma = 0.55$ quando $c^2 = 0.8$. Todos os métodos parecem possuir um crescimento acentuado quando σ é maior que 1.0. O método V tem em média seu vício maior, com algumas oscilações, deixa de ser maior quando $\sigma = 1.0$ em $c^2 = 0.8$ e 0.9 e quando $\sigma = 0.89$ em $c^2 = 0.95$.

Veremos agora as comparações tendo a variância como medida. O quadro abaixo nos mostra o número de vezes que a variância de um determinado método aparece menor que cada um dos métodos.

Quadro 4.19 Número de vezes em que Variância de cada um dos métodos da linha é menor que os da coluna, com coeficiente de correlação 0.8

σ		I	II	III	IV	V	VI	Total
	I		783	1000	976	637	999	4395
	II	217		1000	427	457	999	3100
0.55	III	0	0		0	0	1	1
	IV	24	573	1000		489	999	3085
	V	363	543	1000	511		1000	3417
	VI	1	1	999	1	0		1002
	I		971	1000	1000	774	1000	4745
	II	29		1000	900	281	1000	3210
0.7	III	0	0		0	0	3	3
	IV	0	100	1000		173	1000	2273
	V	226	719	1000	827		1000	3772
	VI	0	0	997	0	0		997
	I		812	1000	1000	293	1000	4105
	II	188		1000	1000	0	1000	3188
0.89	III	0	0		0	0	5	5
	IV	0	0	1000		0	1000	2000
	V	707	1000	1000	1000		1000	4707
	VI	0	0	995	0	0		995
	I		1000	1000	1000	1000	1000	5000
	II	0		1000	0	0	1000	2000
1.0	III	0	0		0	0	6	6
	IV	0	1000	1000		58	1000	3058
	V	0	1000	1000	942		1000	3942
	VI	0	0	994	0	0		994
	I		411	1000	1000	162	1000	3573
	II	589		1000	989	58	1000	3636
3.0	III	0	0		0	0	0	0
	IV	0	11	1000		24	1000	2035
	V	838	942	1000	976		1000	4756
	VI	0	0	1000	0	0		1000
	I		997	1000	1000	837	1000	4834
	II	3		995	342	0	1000	2340
5.0	III	0	5		2	0	122	129
	IV	0	658	998		0	1000	2656
	V	163	1000	1000	1000		1000	4163
	VI	0	0	878	0	0		878
	Total	3348	12526	29856	14893	5243	24134	

Quadro 4.20 Número de vezes em que Variância de cada um dos métodos da linha é menor que os da coluna. com coeficiente de correlação 0.9

σ		I	II	III	IV	V	VI	Total
	I		531	1000	861	440	1000	3832
	II	469		1000	582	406	1000	3457
0.55	III	0	0		0	0	0	0
	IV	139	418	1000		411	1000	2968
	V	560	594	1000	589		1000	3743
	VI	0	0	1000	0	0		1000
	I		931	1000	929	775	1000	4635
	II	69		1000	333	605	1000	3007
0.7	III	0	0		0	1	5	6
	IV	71	667	1000		626	1000	3364
	V	225	395	999	374		997	2990
	VI	0	0	995	0	3		998
	I		94	1000	1000	81	1000	3175
	II	906		1000	952	73	1000	3931
0.89	III	0	0		0	0	0	0
	IV	0	48	1000		65	1000	2113
	V	919	927	1000	935		1000	4781
	VI	0	0	1000	0	0		1000
	I		1000	1000	1000	1000	1000	5000
	II	0		974	0	992	1000	2966
1.0	III	0	26		0	123	289	438
	IV	0	1000	1000		997	1000	3997
	V	0	8	877	3		1000	1888
	VI	0	0	711	0	0		711
	I		807	1000	1000	653	1000	4460
	II	193		1000	450	538	1000	3181
3.0	III	0	0		0	0	1	1
	IV	0	550	1000		544	1000	3094
	V	347	462	1000	456		1000	3265
	VI	0	0	999	0	0		999
	I		998	1000	999	997	1000	4994
	II	2		1000	7	982	1000	2991
5.0	III	0	0		0	5	9	14
	IV	1	993	1000		989	1000	3983
	V	3	18	995	11		1000	2027
	VI	0	0	991	0	0		991
	Total	3904	10467	29541	10481	11306	24301	

Quadro 4.21 Número de vezes em que Variância de cada um dos métodos da linha é menor que os da coluna, com coeficiente de correlação 0.95

σ		I	II	III	IV	V	VI	TOTAL
	I		348	1000	748	285	1000	3381
	II	652		1000	717	282	1000	3651
0.55	**III**	0	0		0	0	0	0
	IV	252	283	1000		275	1000	2810
	V	715	718	1000	725		999	4157
	VI	0	0	1000	0	1		1001
	I		413	1000	820	360	997	3590
	II	587		1000	651	361	997	3596
0.7	**III**	0	0		0	0	0	0
	IV	180	349	1000		348	997	2874
	V	640	639	1000	652		987	3918
	VI	3	3	1000	3	13		1022
	I		1000	1000	985	1000	1000	4985
	II	0		1000	10	1000	1000	3010
0.89	**III**	0	0		0	12	24	36
	IV	15	990	1000		999	1000	4004
	V	0	0	988	1		1000	1989
	VI	0	0	976	0	0		976
	I		919	1000	1000	414	1000	4333
	II	81		1000	911	314	1000	3306
1.0	**III**	0	0		0	0	0	0
	IV	0	89	1000		185	1000	2274
	V	586	686	1000	815		1000	4087
	VI	0	0	1000	0	0		1000
	I		837	1000	915	753	1000	4505
	II	163		1000	376	702	1000	3241
3.0	**III**	0	0		0	0	0	0
	IV	85	624	1000		658	1000	3367
	V	247	298	1000	342		1000	2887
	VI	0	0	1000	0	0		1000
	I		615	1000	998	446	1000	4059
	II	385		1000	518	280	1000	3183
5.0	**III**	0	0		0	0	0	0
	IV	2	482	1000		369	1000	2853
	V	554	720	1000	631		1000	3905
	VI	0	0	1000	0	0		1000
	Total	5147	10013	29964	11818	9057	24001	

Quadro 4.22 Número de vezes em que Variância de cada um dos métodos da linha é menor que os da coluna, com coeficiente de correlação 0.99

σ		I	II	III	IV	V	VI	
	I		189	1000	690	96	1000	2975
	II	811		1000	784	77	1000	3672
0.55	**III**	0	0		91	0	0	91
	IV	310	216	909		121	836	2392
	V	904	923	1000	879		997	4703
	VI	0	0	1000	164	3		1167
	I		222	1000	686	209	1000	3117
	II	778		1000	742	252	1000	3772
0.7	**III**	0	0		92	0	0	92
	IV	314	258	908		193	823	2496
	V	791	748	1000	807		1000	4346
	VI	0	0	1000	177	0		1177
	I		495	1000	768	540	1000	3803
	II	505		1000	544	543	1000	3592
0.89	**III**	0	0		136	0	0	136
	IV	232	456	864		484	739	2775
	V	460	457	1000	516		1000	3433
	VI	0	0	1000	261	0		1261
	I		261	1000	866	94	1000	3221
	II	739		1000	819	144	1000	3702
1.0	**III**	0	0		112	0	0	112
	IV	134	181	888		70	844	2117
	V	906	856	1000	930		1000	4692
	VI	0	0	1000	156	0		1156
	I		266	1000	688	286	1000	3240
	II	734		1000	723	369	1000	3826
3.0	**III**	0	0		106	0	0	106
	IV	312	277	894		259	813	2555
	V	714	631	1000	741		1000	4086
	VI	0	0	1000	187	0		1187
	I		209	1000	724	107	1000	3040
	II	791		1000	786	76	1000	3653
5.0	**III**	0	0		93	0	0	93
	IV	276	214	907		110	827	2334
	V	893	924	1000	890		999	4706
	VI	0	0	1000	173	1		1174
	Total	10604	7783	29370	15331	4034	22878	

Quando $c^2 = 0.8$ e 0.9 a menor variância se apresenta nos métodos I e V, e quando $c^2 = 0.95$ e 0.99 menor variância se atribui aos métodos V e II. A maior variância é dado ao método III, que era de se esperar, pois seu vício é o menor em todos os casos.

Novamente, no próximo quadro veremos o resumo dos resultados obtidos pelos quadros anteriores. Nele relacionamos a informação da porcentagem total que cada método obteve, considerando todos os casos dos desvios e correlações

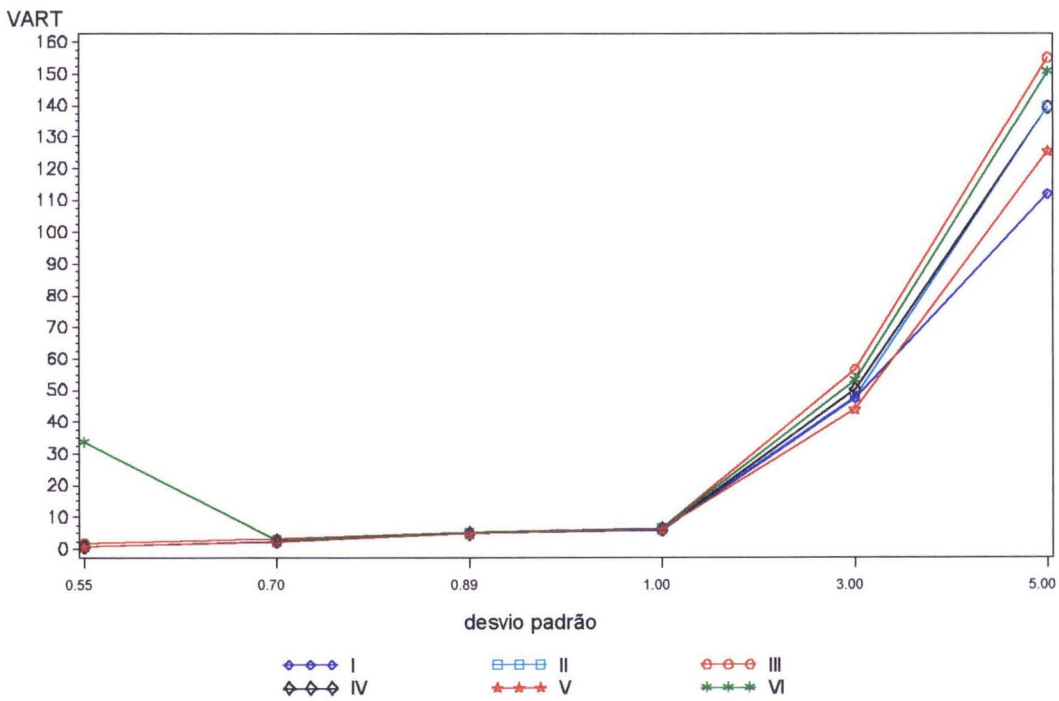
Quadro 4.23 Porcentagem das vezes em que o método teve Variância maior que os outros, considerando os desvios padrões e as correlações teóricas

σ \ c^2		0.8	0.9	0.95	0.99
	I	87.9	76.64	67.62	59.5
	II	62	69.14	73.02	73.44
0.55	III	0.02	0	0	1.82
	IV	61.7	59.36	56.2	47.84
	V	68.34	74.86	83.14	94.06
	VI	20.04	20	20.02	23.34
	I	94.9	92.7	71.8	62.34
	II	64.2	60.14	71.92	75.44
	III	0.06	0.12	0	1.84
0.7	IV	45.46	67.28	57.48	49.92
	V	75.44	59.8	78.36	86.92
	VI	19.94	19.96	20.44	23.54
	I	82.1	63.5	99.7	76.06
	II	63.76	78.62	60.2	71.84
	III	0.1	0	0.72	2.72
0.89	IV	40	42.26	80.08	55.5
	V	94.14	95.62	39.78	68.66
	VI	19.9	20	19.52	25.22
	I	100	100	86.66	64.42
	II	40	59.32	66.12	74.04
	III	0.12	8.76	0	2.24
1.0	IV	61.16	79.94	45.48	42.34
	V	78.84	37.76	81.74	93.84
	VI	19.88	14.22	20	23.12
	I	71.46	89.2	90.1	64.8
	II	72.72	63.62	64.82	76.52
3.0	III	0	0.02	0	2.12
	IV	40.7	61.88	67.34	51.1
	V	95.12	65.3	57.74	81.72
	VI	20	19.98	20	23.74

	I	96.68	99.88	81.18	60.8
	II	46.8	59.82	63.66	73.06
5.0	**III**	2.58	0.28	0	1.86
	IV	53.12	79.66	57.06	46.68
	V	83.26	40.54	78.1	94.12
	VI	17.56	19.82	20	23.48

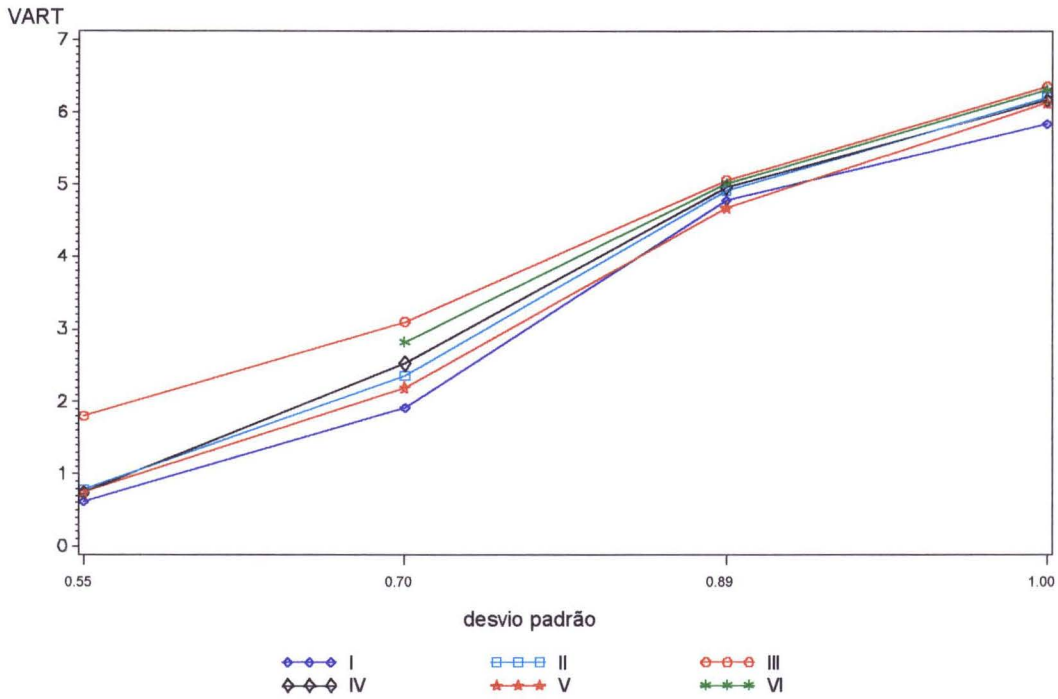
Veremos agora os gráficos da variância como função do desvio padrão para cada uma das diferentes correlações. Novamente, devido ao crescimento brusco dos métodos quando $\sigma > 1$. Faremos para cada caso dos coeficientes de correlação um segundo gráfico onde ampliaremos o intervalo (0.55, 1.0) do desvio padrão.

Gráfico 4.19 Variância total em função do desvio padrão com $c^2 = 0.8$



Pelo mesmo motivo da variável vício, também faremos dois gráfico para cada uma das correlações. Assim, traçaremos os gráficos de numeração par como 4.20 para uma melhor visualização dos gráfico de numeração ímpar.

Gráfico 4.20 Ampliação do intervalo (0.55, 1.0) com $c^2 = 0.8$



No caso de correlação 0.8, observamos, graficamente, que o método I é menor quando σ = 0.55, 0.7, 1.0 e 5.0. Nos outros desvios o método V apresenta a menor variância. Os métodos com as maiores variâncias são III e VI.

Gráfico 4.21 Variância total em função do desvio padrão com $c^2 = 0.9$

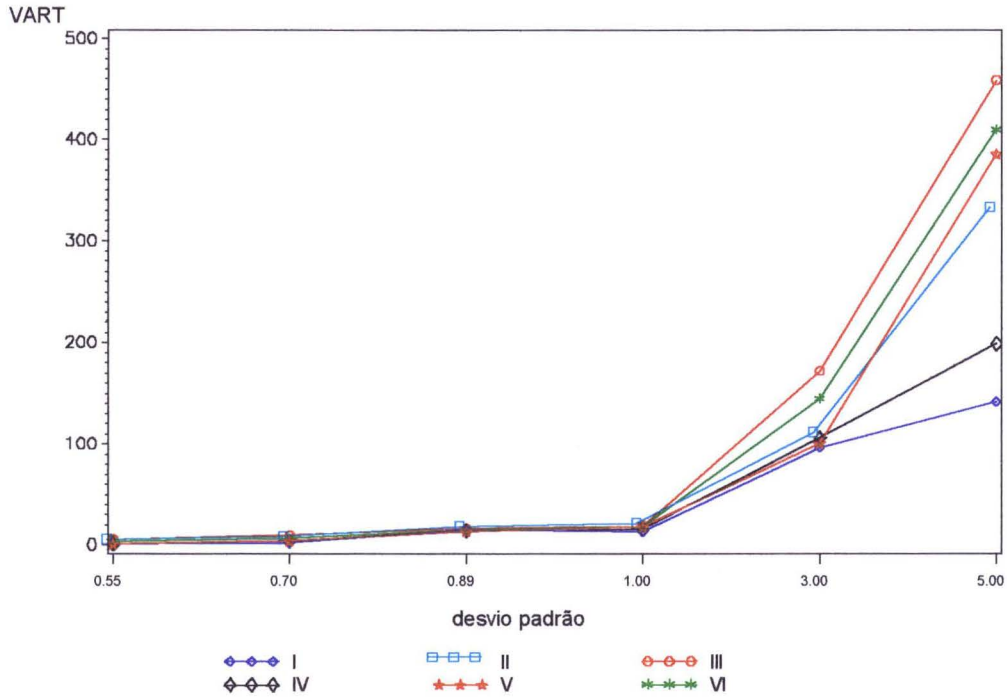
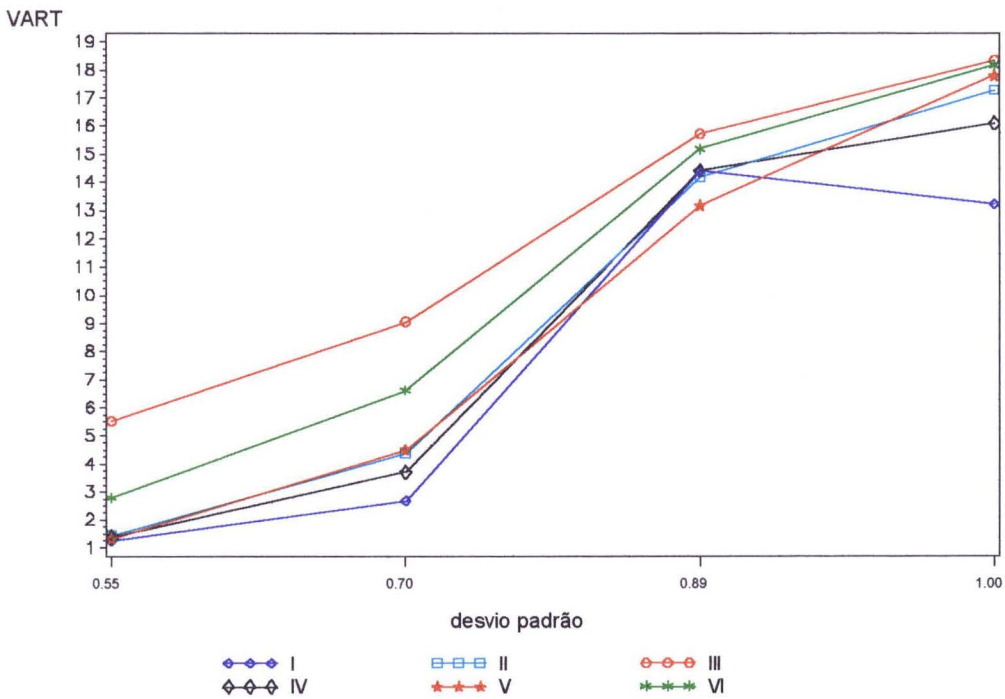


Gráfico 4.22 Ampliação do intervalo (0.55, 1.0) com $c^2 = 0.9$



Neste segundo caso da correlação 0.9, I sempre tem a menor variância com exceção de σ=0.89, novamente, atribuindo ao método V a menor variância neste desvio. Já o de maior variância se atribui somente ao método III.

Gráfico 4.23 Variância total em função do desvio padrão com $c^2 = 0.95$

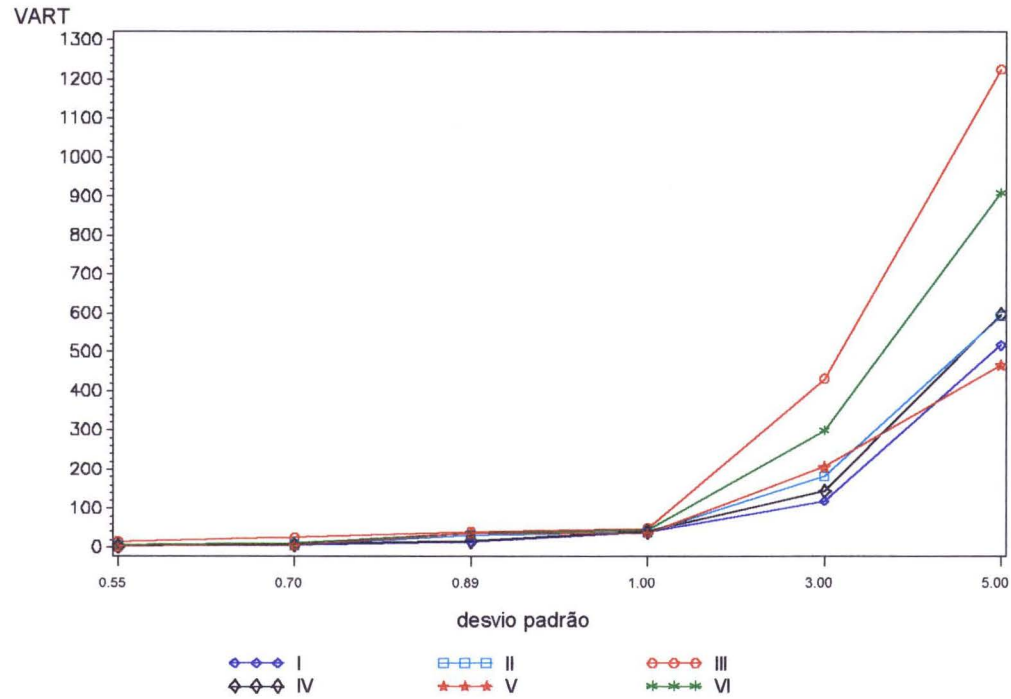
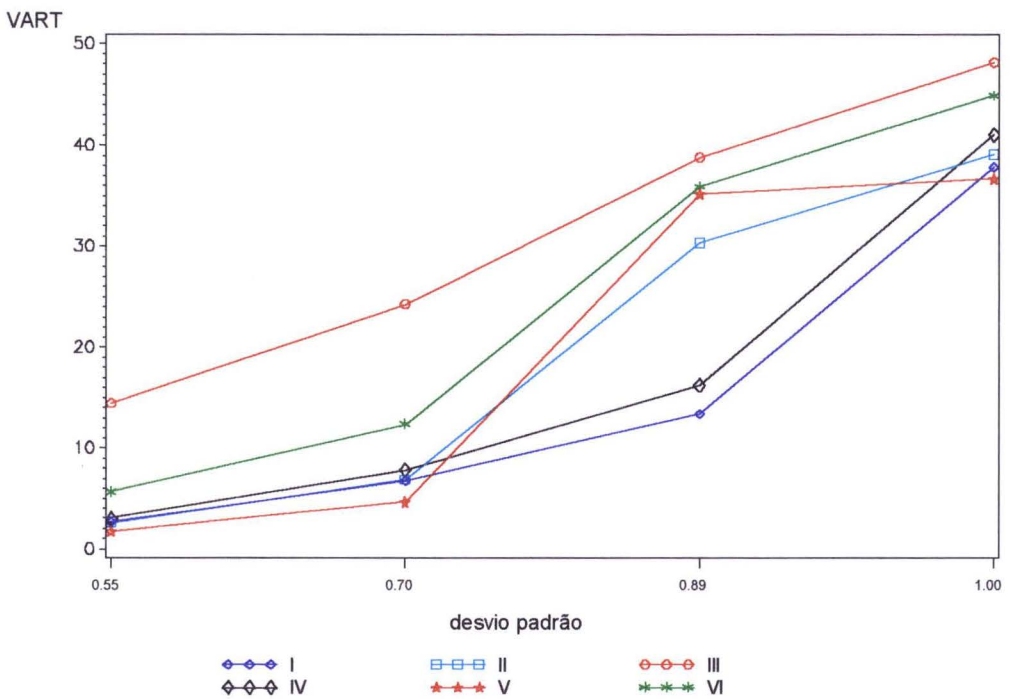


Gráfico 4.24 Ampliação do intervalo (0.55, 1.0) com $c^2 = 0.95$



Nos gráficos 4.23 e 4.24, consideramos a correlação teórica 0.95. Vemos uma certa oscilação entre os métodos I e V, na posição de menor variância, sendo que o método I se apresenta menor somente quando σ = 0.89 e 3.0 e o método V nos demais desvios. Novamente, III possui maior variância , em todas situações dos desvios padrões.

Gráfico 4.25 Variância total em função do desvio padrão com $c^2 = 0.99$

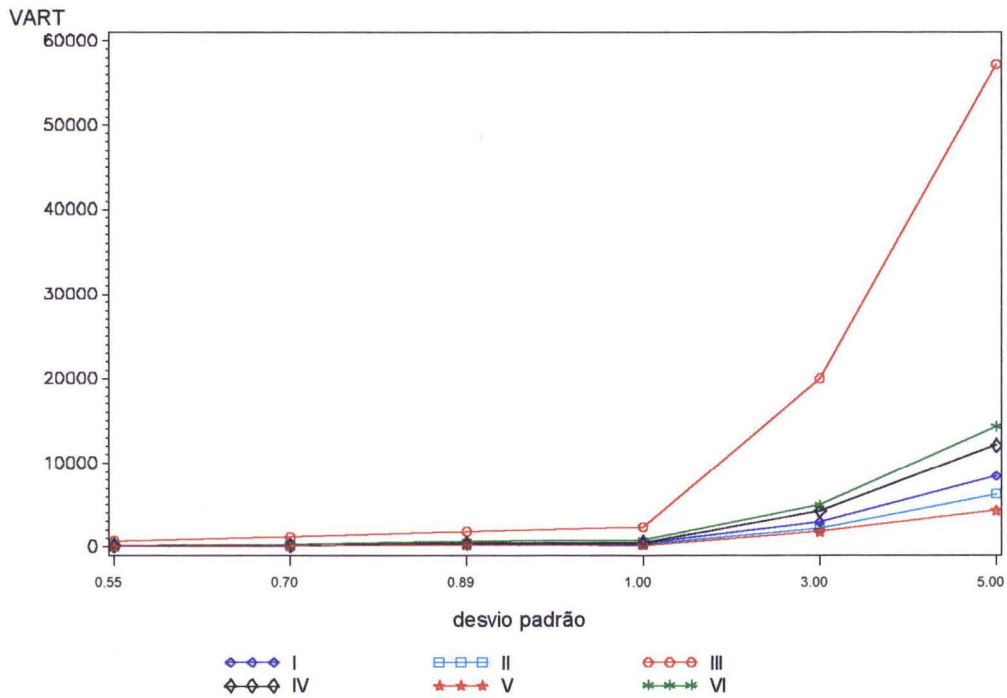
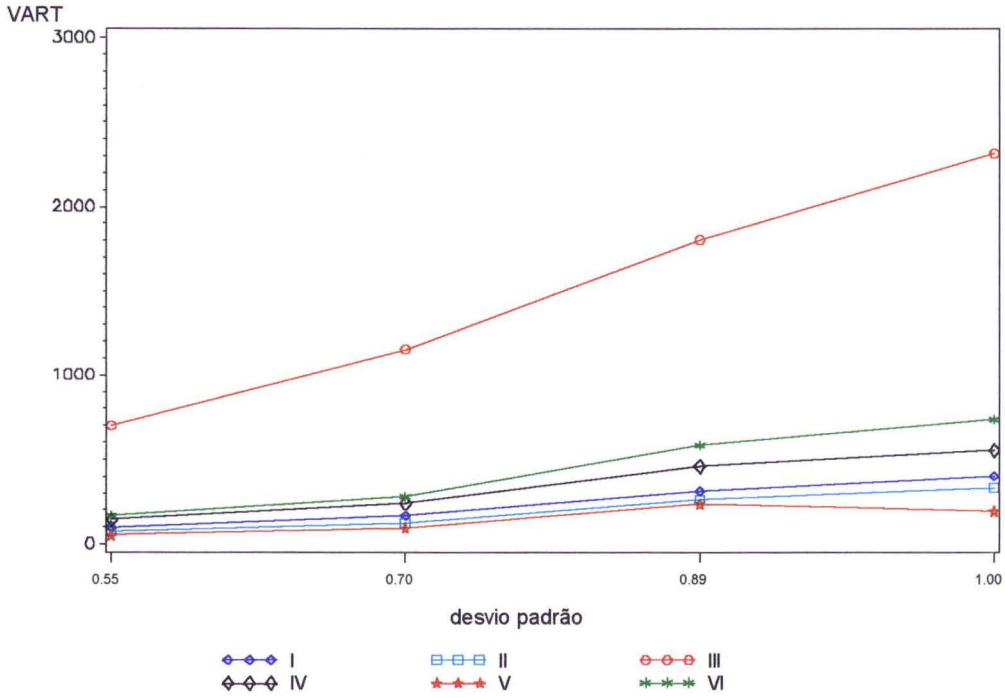


Gráfico 4.26 Ampliação do intervalo (0.55, 1.0) com $c^2 = 0.99$



Nestes gráficos de correlação teórica 0.99, o método que se apresenta menor é o V, depois deste, com a segunda menor variância temos o método II. Novamente III tem a maior variância, o segundo maior valor é dado ao método VI.

4.2 Exemplo

Neste exemplo[2], estamos interessados em estudar o índice de fundo que um banco detém em seu poder, isto é, a valorização ou desvalorização de suas ações aplicadas nas duas bolsas de valores, BOVESPA e BVRJ. Os dados estão na tabela abaixo.

[2] O programa do exemplo consta no apêndice deste trabalho.

Tabela Índice mensal das bolsas BOVESPA e BVRJ e de um determinado banco.

Obs	y	x_1	x_2
1	-10.9461	-13.0410	-11.1688
2	-1.6413	-17.7506	-17.3155
3	-0.6031	-11.4137	-10.2064
4	10.5693	23.1235	20.2667
5	-2.3333	-5.9790	-3.6763
6	-0.3675	-6.3347	-5.4657
7	-0.9236	3.9580	1.5315
8	-1.9104	7.8063	6.0877
9	0.3133	5.7515	4.8854
10	-12.2864	-13.4722	-12.3083
11	1.9216	4.0361	0.4913
12	-0.7256	-4.0455	-1.8014
13	11.2314	17.7588	15.6853
14	3.2779	-5.1534	-3.4606
15	1.8832	-1.3566	-1.6117
16	0.4814	3.0239	1.4809
17	7.4388	9.7195	9.5222
18	8.6665	4.3537	3.6327
19	4.13196	0.2213	-1.7669
20	-0.87907	1.0821	1.9049
21	-1.10408	1.8028	0.9411
22	6.91682	0.0915	1.7933

Fonte: Banco Central.

Assim construímos o seguinte modelo:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

onde, x_1: corresponde aos índices mensais dos negócios realizados na bolsa de São Paulo, no período de janeiro de 95 a outubro de 96, x_2: relativo aos índices mensais dos negócios realizados na bolsa do Rio de Janeiro, y: refere-se ao rendimento de um determinado banco e ε é a variável aleatória que corresponde ao erro diversificado.

Construído o modelo devemos estimá-lo, isto é, desejamos saber qual o risco que se corre em aplicar no fundo.

Como os índices mensais das bolsas estão correlacionados, a matriz X, por sua vez, apresentam-se mal condicionada. Com isso, o que se constuma fazer é eliminar uma das variáveis. Nessas situações os economistas eliminam a variável que corresponde à bolsa do Rio de Janeiro, isto porque esta é, relativamente, menor que a bolsa de São Paulo em volume de negócios. Obtendo assim, o rendimento das cotas do banco em função apenas da bolsa BOVESPA. Contudo, existem algumas ações, por exemplo, da PETROBAS, TELEBRAS, das estatais em geral, que são muito negociadas no Rio e com sua eliminação

estamos desprezando informações importantes considerando que o banco também negocia com estas ações na BVRJ.

Assim, tendo que as duas variáveis são importantes ao modelo, sugerimos o método "ridge" para se obterem os estimadores, visto que as regressoras apresentam-se mal condicionadas. Desta forma, poderemos analisar o rendimento das cotas do banco sem perda de informações.

Lembramos que os dados foram centrados e padronizados, como definido no primeiro capítulo. Baseado no programa feito neste capítulo, calculamos os diagnósticos de multicolinearidade, o EQMT(0), EQMTP(0) e os estimadores de mínimos quadrados.

Quadro 4.24: Diagnóstico de multicolinearidade

η	VIF
200.92317	50.732036

$$EQMT(0) = \sigma^2 \sum_{i=1}^{2} \frac{1}{\lambda_i} = 1667.1648$$

$$EQMTP(0) = 32.86217$$

$$\mathbf{b} = \begin{pmatrix} 1.0505318 \\ -6.776629 \\ 26.724072 \end{pmatrix},$$

a variância de cada um dos estimadores acima são :

$$Var(\mathbf{b}) = \begin{pmatrix} 0.7468675 \\ 8.2564314 \\ 1658.9083 \end{pmatrix}$$

Alguns dos resultados podem ser confirmados pelo PROC REG do SAS.

Quadro 4.25 : Análise da Variância

```
Model: MODEL1
Dependent Variable: Y3

                          Analysis of Variance

                                Sum of          Mean
          Source          DF    Squares        Square      F Value       Prob>F

          Model            2    401.48814    200.74407       12.217       0.0004
          Error           19    312.19045     16.43108
          C Total         21    713.67859

              Root MSE        4.05353     R-square        0.5626
              Dep Mean        1.05053     Adj R-sq        0.5165
              C.V.          385.85470


                          Parameter Estimates

                    Parameter       Standard    T for H0:                    Variance
  Variable  DF      Estimate          Error   Parameter=0    Prob > |T|     Inflation

  INTERCEP   1      1.050532     0.86421473         1.216        0.2390    0.00000000
  X1         1     -6.776721    28.87177911        -0.235        0.8169    50.73192240
  X2         1     26.724157    28.87177445         0.926        0.3663    50.73192240
```

Observamos na análise de variância que o modelo é significativo, mas quando vemos a estimativa dos parâmetros obtemos que nem o índice BOVESPA e nem BVRJ são significativos e que o fator de inflação da variância é muito grande. Isto é uma consequência do mal condicionamento das regressoras. Nestas condições podemos utilizar a regressão "ridge" para obtermos os estimadores.

Escolhemos alguns dos métodos utilizados na simulação para aplicarmos neste exemplo. Os métodos escolhidos foram: I e V. Estes foram escolhidos por apresentarem o EQMT e a variância, em algumas situações, melhores que os demais. Vejamos qual os resultados obtidos. Para tal, calculamos o EQMTP, o EQMT, o vício e a variância e os respectivos estimadores "ridge".

Quadro 4.26: Resultados obtidos dos métodos "ridge"

Métodos	**I**	**V**
Variância	156.17902	7.4332505
Vício	350.06507	674.54846
EQMT	506.24409	681.98171
EQMTP	18.872039	20.620934

No quadro 4.26 podemos ver a vantagem do método "ridge" sobre os mínimos quadrados, pois apresentam-se com variância, EQMT e EQMTP menores. O método V apresentou uma variância, substancialmente, menor, com uma diferença maior que 1660 dos mínimos quadrados. O método I também nos forneceu a variância, relativamente, pequena com diferença maior que 1511. O EQMTP é o menor nos dois casos. Quanto ao EQMT o método I possui o menor valor entre eles. Desses resultados, podemos confirmar as vantagens do método "ridge", nos casos de mal condicionamento.

Quanto aos estimadores "ridge", em ambos os casos, o intercepto é igual a $\bar{y}$ = 1.0505318 e os estimadores de β_1 e β_2 são mostrados nos vetores:

$$\mathbf{b_I(k)} = \begin{pmatrix} -5.743929 \\ 8.0425678 \end{pmatrix}$$

$$\mathbf{b_V(k)} = \begin{pmatrix} -5.803378 \\ 0.770243 \end{pmatrix}$$

Observamos que o resultado do coeficiente que corresponde ao BVRJ, usando o método de mínimos quadrados é muito maior que dos métodos "ridge". Essa diferença é, justamente, atribuída ao coeficiente onde sua variância é muito grande. Com a regressão "ridge" esse valor diminui consideravelmente. Os métodos I e V também nos mostra uma diferença nesse coeficiente. Essa diferença, possivelmente, pode ser explicada pelo vício, visto que o vício do método V é o dobro do I. Diante desses resultados, comprovamos, neste exemplo, as vantagens dos estimadores ridge no que se refere a variância menor.

4.3 Conclusão

4.3.1 Retrospectiva dos Resultados

O objetivo deste trabalho desde o início era comparar os diferentes métodos da regressão "ridge" e mostrar suas vantagens sobre os estimadores de mínimos quadrados,

quando os dados são mal condicionados. Na tentativa de mostrar ao leitor as vantagens de cada método proposto, simulamos um conjunto de dados como foi mostrado no capítulo em questão fizemos várias comparações e análises. Estas tinham como fundamentação os dois primeiros capítulos, onde pudemos ver a base teórica da regressão múltipla e multicolinearidade, para a compreensão do terceiro capítulo que trata da regressão "ridge".

Assim, todas as informações foram usadas neste capítulo sem muitas citações. Como, por exemplo, no cálculo das medidas de multicolinearidade.

Vimos que por estas análises tivemos a matriz, no caso de correlação 0.95 e 0.99, com mal condicionamento evidenciado. Incluindo essa informação analisamos todas as variáveis de interesse.

Na análise da variável EQMTP, observamos que, exceto o método V, todos demais possuem os valores dos EQMTP(k) menores que EQMTP(0). Em particular, neste caso, o método que se apresentou com menor EQMTP(k) foi o I.

Quanto ao vício vimos que em todos os casos dos coeficientes de correlações e dos desvios padrões temos unanimidade nos resultados a favor do método III. Este em todos os casos apresentou-se melhor que o método de mínimos quadrados. No entanto, os valores do EQMT e EQMTP, na maioria das vezes, foi maior que dos outros métodos.

Observamos em relação a variância que tanto o método I como o V nos fornecem bons estimadores quando se deseja variância pequena, não se importando com o tamanho do vício. Dentre os dois métodos, no caso da correlação 0.99, o estimador que melhor se ajustou foi o obtido pelo método V, nos demais casos pelo método I. Vimos, também, no capítulo 3 que quanto menor a variância maior é o vício do estimador, contudo, notamos que, graficamente, o método V quase sempre nos fornece um vício maior que de todos os outros métodos. Algumas vezes, esta posição é atribuída ao método I. Pesando cada informação, o leitor pode escolher entre estes o método para obtenção do seu estimador "ridge", quando se deseja uma variância muito pequena.

Na variável EQMT(k), os métodos que se destacaram foram I e IV. Sendo que o I, obteve 100% dos casos seu EQMT(k) menor que de todos outros métodos. Este é um importante fator, na escolha do método a utilizar. Neste caso, consideramos também o fato da dificuldade de se obter o método. Por exemplo, o método I é ótimo no sentido que esse minimiza o EQMT e nos fornece em grande parte uma variância menor que dos outros métodos. Mas este método exige que obtenhamos uma matriz diagonal K com diferentes valores adicionadas na diagonal da matriz $W^{T}W$ e não um único valor de k.

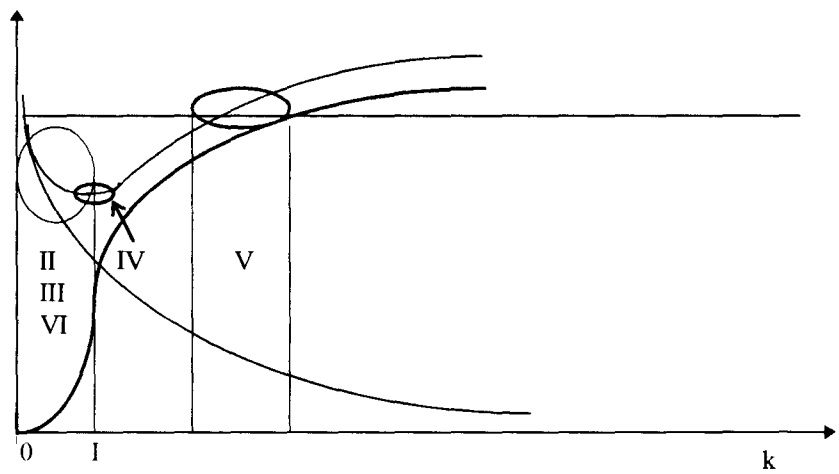
Desta forma para os leitores que de uma forma mais rápida deseja obter um único valor de k que nos forneça um EQMT tão pequeno quanto do método I, sugerimos segundo nossos resultados das simulações o método IV, que como pudemos observar, graficamente, obtém depois do método I o menor EQMT em todos os casos dos desvios padrões e correlações.

Na seção 4.2, vimos o exemplo das bolsas de valores. Nele, obtivemos o método V com a menor variância, e o método I recebeu os demais atributos.

4.3.2 Aspectos Principais dos Métodos

No capítulo 3 mostramos o gráfico dos valores de k como função do EQMT, da variância e do vício. Apresentaremos, agora, segundo os resultados das simulações a localização de cada um dos métodos propostos na projeção do gráfico. As elipse na figura 4.1 representam a região onde se localiza cada método.

Figura 4.1



O método I, como foi visto na teoria do capítulo 3 e obtido nos resultados da simulação é exatamente o ponto de mínimo global da função EQMT. O método III foi proposto com o intuito de se obter uma resolução aproximada do método interativo de Hoerl utilizando apenas um critério de convergência/divergência. Entretanto, obtivemos

como o método que minimiza o vício. Os outros métodos que também nos fornecem vícios pequenos são II e VI, localizados entre zero e o do método I. O método IV, teoricamente, nos fornece o valor de mínimo local da função EQMT de forma que minimiza o vício. Retratamos, segundo nossa simulação, que este método, realmente, nos fornece um valor razoável do EQMT e um vício não muito grande, de forma que confirmamos as pesquisas bibliográficas literárias. O método V retrata uma resolução Bayeseana para o problema. Em nossas simulações, o método em questão nem sempre nos fornece EQMT(k) e EQMTP(k) menor EQMT(0), localizado como mostra o gráfico 4.3. O método em grande parte nos fornece valor pequeno da variância, mostrando-nos, na seção 4.2, uma variância, substancialmente, menor que dos mínimos quadrados e, também, do método I.

4.3.3 Conclusão Geral

De uma forma geral vimos que, baseados no vício, variância e erro quadrático médio do estimador e do predito os métodos III, I e V, I e IV, e I foram os que melhor se apresentaram, respectivamente. Vimos, também, que todos os métodos se apresentaram melhor que os estimadores de mínimos quadrados, exceto o V na ausência do mal condicionamento forte, possui frequência maior que zero de M>1. Considerando, conjuntamente, todas essas informações e a presença do mal condicionamento, supondo que não podemos eliminar variáveis do modelo, os métodos que melhor ajustam a cada um dos casos podem ser destacados no seguinte quadro.

Variáveis	Métodos
EQMTP	I
Variância	V
Vício	III
EQMT	I

Diante disso, concluímos que a regressão "ridge" nos fornece, apesar de viciados, estimadores mais precisos. Esses resultados foram obtidos quando os dados

estavam com mal condicionamento mais acentuado como também nos casos das menores correlações. Assim, incentivo o emprego dos estimadores "ridge" quando há suspeitas do mal condicionamento da matriz. Além disso, sugiro o estudo da análise de variância e intervalos de confiança para estes estimadores, como um estudo mais detalhado da regressão "ridge".

Apêndice

Apêndice A

Este apêndice apresenta a geração dos dados que forma a matriz das regressoras. Esta geração refere-se ao capítulo 4.

```
libname dado 'c:\users\acris\programa';
%Macro data;                          /* Esta Macro data gera */
 data dado.a;                         /*números speudo-aleatórios */
 %do n=1 %to 4;                       /* normal (0,1)*/
  %do i=1 %to 15;
    z1=rannor(42495+34*&i+36*&n);
    z2=rannor(53456+66*&i+85*&n);
    z3=rannor(52848+88*&i+54*&n);
    z4=rannor(16695+91*&i+93*&n);
    output;
  %end;
 %end;
 run;
%Mend;
%data;

%Macro data1;                         /*Esta Macro data1 atribui aos*/
data _null_;                          /*valores gerados em data os */
 set dado.a end=eof;                  /*nomes z1,...,z4/*
  call symput('z1'||left(_n_),z1);
  call symput('z2'||left(_n_),z2);
  call symput('z3'||left(_n_),z3);
  call symput('z4'||left(_n_),z4);
 run;
%Mend;
%data1;
```

Estas macros são uma espécie de transporte para comunicação no proc iml.

Apêndice B

Este apêndice, apresenta o programa do PROC IML - SAS relativo ao exemplo dado no capítulo 4. A versão do *software The SAS System* utilizada é a 6.03.

Programa do exemplo.

```
Proc Iml;
/* ---------- Dados do Exemplo ---------- */
x2={-13.041,-17.7506,-11.4147,23.1235,-5.9790,-6.3347,3.9580,7.8063,5.7515,-
13.4722,4.0361,-4.0455,17.7588,-5.1534,-1.3566,3.0239,
9.7195,4.3537,0.2213,1.0821,1.8028,0.0915};
x3={-11.1688,-17.3155,-10.2064,20.2667,-3.6763,-5.4657,1.5315,6.0877,4.8854,-
12.3083,0.4913,-1.8014,15.6853,-3.4306,-1.6117,1.4809,
9.5222,3.6327,-0.7669,1.9049,0.9411,1.7933};
Y3={-10.9461,-1.6413,-0.6031,10.5693,-2.3333,-0.3675,-0.9236,-1.9104,0.3133,-
12.2864,1.9216,-0.7256,11.2314,3.2779,1.8832,0.4814,
7.4388,8.6665,4.1320,-0.8791,-1.1041,6.9168};
M                                   /* Neste espaço todas as matrizes devem */
                                    /*ser inicializadas*/
/* ----------Padronização das matrizes---------- */
s[1]=sum(x2); sq[1]=ssq(x2);
s[2]=sum(x3); sq[2]=ssq(x3);

do i=1 to 2;
  u[i]=s[i]/22;                     /*u é média e v é (n-1)variância*/
  v[i]=(sq[i]-22*u[i]##2);
end;
```

```
do k=1 to 22;
   x[k,1]=1;
   x[k,2]=(x2[k]-u[1])/sqrt(v[1]);
   x[k,3]=(x3[k]-u[2])/sqrt(v[2]);                    /*padronização da matriz X*/
  end;

a=x`*x;                                               /* a é a matriz X'X */

  call svd (w,q,p,a);                                 /*decomposição de valores*/
                                                      /*singulares, retorna os auto-*/
                                                      /*-valores, q,  e autovetores, w.*/

/* cálcula o estimador de mínimos quadrados, quadrado médio do erro, as somas de quadrados*/
    c=inv(a);                                         /* inversa da matriz a=X^TX*/
   ymedio=sum(y3)/22;
    alfaest=c*x`*y3;
   yest=x*alfaest;
   resid=y3-yest;
   sse=resid`*resid;                                  /*SQE*/
   regres=yest-ymedio*ymed;
   total=y3-ymedio*ymed;
   sst=ssq(total);                                    /*SQT*/
   ssr=ssq(regres);                                   /*SQR*/
   qmr=ssr/2;                                         /*QMR*/
   qme=sse/19;                                        /*QME*/
   print sst ssr sse qme qmr;                         /*imprime as variáveis indicadas*/
```

```
/*---------- obtemos índice de condição e o VIF---------- */

nq[1]=q[2];                          /*nq é o vetor de autovalores da
nq[2]=q[3];                          /*matriz X sem o intercepto*/

do i=1 to 3;
      IC[i]=nq[1]/nq[i];             /*  índice de condição */
end;
 VIF=c[2,2];                         /* fator de inflação da variância*/
  R2 =ssr/sst;                       /*coeficiente de determinação*/
  print R2  VIF IC;
 do i=1 to 2;                        /*os alfaes são os EMQ[1] da matriz X */
    alfaes[i]=alfaest[i+1];          /*sem o intercepto*/
  end;

/* ---------- Cálcula os valores de k para os métodos I e V, usados no exemplo. ----------*/

kI=qme/alfaes##2;
aux=nq#alfaes##2;
kV=2*qme/sum(aux);

   do i=1 to 2;
     kI[i]=qme/alfaes [i]**2;
   end;
```


[1] EMQ é abreviatura de estimador de mínimos quadrados

```
/* ------------------------------Estimadores Ridge ---------------------------------------- */
  do i=1 to 3;
        ridV[i]=nq[i]*alfaes[i]/(nq[i]+kV);
        ridI[i]=nq[i]*alfaes[i]/(nq[i]+kI[i]);
  end;

  do i=1 to 3;                                  /*m cálcula EQMP de cada estimador*/
    m[i]=qme;
    mV[i]=nq[i]*(q[i]*qme+(kV*alfaes[i])**2)/(nq[i]+kV)**2;
    mI[i]=nq[i]*(q[i]*qme+(kI[i]*alfaes[i])**2)/(nq[i]+kI[i])**2;
end;
  pred=SUM(M);                                  /* pred é o EQMTP*/
   predV=sum(mV);
   predI=sum(mI);
  do i=1 to 3;                                  /* var é a variância de cada estimador*/
    var[i]=qme/nq[i];
    varI[i]=qme*nq[i]/(nq[i]+kI[i])**2;
    varV[i]=qme*nq[i]/(nq[i]+kV)**2;
  end;
  varemq=sum(var);                              /* varemq é a variância total dos EMQ*/
  variaV=sum(varV);                             /* varia é a variância total*/
  variaI=sum(varI);
  do i=1 to 3;                                  /* mest é o vício de estimador*/
   mestI[i]=(kI[i]*alfaest[i])**2/(nq[i]+kI[i])**2;
   mestV[i]=(kV*alfaest[i])**2/(nq[i]+kV)**2;
  end;
```

```
    vcV=sum(mestV);                    /* vc é o vício total*/
    vcI=sum(mestI);
   qmestV=vcV+variaV;                  /* qmest é o EQMT*/
   qmestI=vcI+variaI;

/*---------Imprime, na ordem, ridge, variância total , vício total, EQMT e EQMTP -------*/
     print alfaest ridI ridV;
     print varemq variaI variaV;
     print vcI vcV;
     print varemq qmestI qmestV;
     print pred predI predV;
quit;
```

Bibliografia

BERK, K. N., (1997). Tolerance and Codition in Regression Computations. *Journal of the American Statistical Association*, 72, 360, 863 - 866.

BESLEY, D. A. (1991). *Conditionng Diagnostics Collinearity and Weak Data in Regression*. 1.ed. New York : John Wiley & Sons, Inc.

BIRKES, D. e DODGE, Y. (1993). *Alternative Methods of Regression*. 1.ed. New York : John Wiley & Sons, Inc.

DRAPER, N. R. e SMITH, H. (1981). *Applied Regression Analysis*. 2.ed. New York :John Wiley & Sons.

GIBBONS, D. G. (1981). A Simulation Study of Some Ridge Estimators. *Journal of the American Statistical Association*, Warren, 96, 373, 131 - 139.

GRAYBILL, F. A. (1983). *Matrices With Applications in Statistics*. 2.ed. Belmont, Caliif.:Wadsworth, inc.

HEMMERLE, W.J. (1975). An Explicit Solution for Generalized Ridge. *Technometrics*, Island, 17, 3, 309- 314.

HOERL, A. E. e KENNARD, R. W. (1970). Ridge Regression: Biased Estimation or Nonorthogonal Problems, *Technometics*, 12,1, 55 - 67.

_____.(1976). Ridge Regression: Iterative Estimation of the Biasing Parameter. *Communication in Statistics - Theory and Methods*, Delaware, A5(1), 77 - 88.

_____ e BALDWIN, K. F. (1975). Ridge Regression: Some Simulations. *Communication in Statistics*, 4(2), 105 - 123.

_____ e SCHERENEMEYER, J e Hoerl, R.W. (1986). *A Simulation of Biased Estimation and Subset Selection Regression Techniques*. Technometrics, 8, 4, 369 - 380.

JUDGE, G.G et al. (1986). *The Theory and Practice of Econometrics*. 1.ed. New York :John Wiley & Sons, inc.

LAWLESS, J. F. (1981). Mean Squared Error Properties of Generalized Ridge Estimators. *Journal of the American Statistical Association*, Ontário, 76,374, 462 - 466.

_____.(1978). Ridge and Related Estimation Procedures: Theory and Practice. *Communication in Statistics - Theory and Methods*, A7(2), 139 - 164.

_____ e WANG, P. (1976). A Simulations Study of Ridge and Other Regression Estimators, *Communication in Statistics - Theory and Methods*, A5(4), 307 - 323.

LEE, Tze-San e CAMPBELL, D. B. (1985). Selectiong the Optimum k in Ridge Regression, *Communication in Statistics* - Theory and Methods, 14(7). 1589 - 1604.

MALLOWS, C. L. (1973). Some Comments em C_p.. *Technometrics*, New Jersey, 15, 4, 661 - 675.

MCDONALD, G. C. (1980). Some Algebraic Properties of Ridge Coefficients. *J.R.Statistical Soc. B*, Michigan, 42,1, 31 - 34.

_____ e GALARNEAU, D. I. (1975). A Monte Carlo Evaluation of Some Ridge Type Estimators. *Journal of the American Statistical Association*, 70, 350, 407 - 415.

MONTGOMERY, D. C e PECK, E. A. (1992). *Introduction to Linear Regression Analysis*, 2.ed. New York :John Wiley & Sons, Inc.

PEARSON, E. S. e KENDALL, M. G.(1820). *Studies in the History of Statistics and Probability,* 1.ed. London, Charles Griffin & Company Limited.

RAO, C. R. (1973). *Linear Statistical Inference and its Applications.* 2.ed. New York :John Wiley & Sons.

RILEY, J. D. (1995). Solving Systems of Linear Equations With a Positive Definite Symmetric, but possibly ill-conditioned matrix. *Matematics of Computation,* 9, 96 - 101,

SAS/IMLTH User's Guide, Release 6.03 Edition., Cary, NC, USA. SAS Institute Inc(1988).

STEWART, G. W. (1987). Collinearity and Least Squares Regression. *Statistical Science*, 2,1,68 - 100.

THEIL, H. (1971). *Principles of Econometrics.* 1.ed. New York :John Wiley & Sons, Inc.,

WEISBERG, S. (1985). *Applied Linear Regression*. 2.ed. New York :John Wiley & Sons, Inc.,

WETHERILL, B. G. et al. (1986). *Regression Analysis With Applications*. London: Chapman and Hall Ltd.

Concordo com a reprodução desta dissertação.
Campinas, 17 de novembro de 1997.